In re Application of: Pietrokovski et al
Serial No.: 10/534,544
Filed: May 10, 2005
Office Action Mailing Date: Dec 18, 2008

Examiner: Ogunbiyi
Group Art Unit: 1645
Attorney Docket: 29489

REMARKS

Reconsideration of the above-identified application in view of the amendments above and the remarks following is respectfully requested.

Claims 1 and 5-121 are in this Application. Claims 19-121 have been withdrawn from consideration. Claims 1 and 5-18 have been rejected. Claims 10 and 16 have now been amended.

Rejections Maintained

The Examiner has maintained the rejection of claims 16-18 under USC 112, Second Paragraph.

The Examiner states that these claims are vague and indefinite as to which part of a virus or a cell the affinity tag binds to.

The instant specification provides several specific examples of affinity tags that can be used by the present invention, including "streptavidin, His-tags, strep-tags, epitope tags, maltose-binding proteins, and chitin-binding domains." Such examples clearly demonstrate the type of tags that can be used with the present invention and clearly identify, to the ordinary skilled artisan, the part or parts of the virus or cell that would be bound by such tags.

Detailed description regarding use of such affinity tags is provided in section [0109]-[0118] of the published application.

In addition, it should be noted that affinity tags are well known in the art and are routinely used for affinity-based purification. The present invention does not teach or suggest novel affinity tags or novel uses for such tags, but rather describes art-acceptable use of affinity tags in combination with the autoprocessing segment of the present invention. In light of the description and examples provided in the instant specification and the wealth of knowledge available in the art, Applicant is of the opinion that one of ordinary skill in the art would be more than capable of modifying the present invention to include affinity tags suitable for purification of any cell or virus desired.

Notwithstanding from the above and in the interest of expediting prosecution of this case, Applicant has elected to amend claim 16 to recite "wherein said molecule is displayed on a virus or a cell". Support for such an amendment is provided throughout the instant specification, see for example, sections [0172]-[0174].

New Rejections

35 U.S.C. § 112, First Paragraph Rejections

The Examiner has rejected claim 16 under 35 USC 112, First Paragraph, as failing to comply with the written description requirement.

The Examiner states that the specification does provide support for the binding of a molecule that forms a part of a virus or cell.

Claim 16 has now been amended to clarify that said molecule is displayed on the virus or cell. Support is provided by sections [0172]-[0174] of the published application.

The Examiner also rejected claims 1 and 5-18 under 35 USC 112, First Paragraph, as failing to comply with the written description requirement.

The Examiner states that the sequence set forth by SEQ ID NO: 31 encompasses an extremely large number of different species because of the variability allowed in the sequence.

The Examiner further states that the specification teaches in Example 4 page 56 that a chimeric protein which comprises the type A BIL domain BIL4_cloth (SEQ ID NO:31) has the capacity to efficiently display auto splicing and carboxy terminal auto-cleaving, but does not set forth which variant was used for the experiment, i.e. its specific sequence.

The Examiner fails to understand that SEQ ID NO:31 embodies a specific sequence of one type of an autoprocessing segment.

By aligning numerous sequences, the present inventor clearly demonstrated that BIL domains share consensus sequence motifs that are spaced apart by regions of great variance in amino acid sequences (see Figures 3a-g).

In SEQ ID NO:31 (as with any BIL domain) intervening sequences (designated by Xaa in SEQ ID NO:31) are not a part of the autoprocessing-defining sequence. These sequences can include any amino acid sequence, since their only function is to space apart the sequences that define autoprocessing. Since the sequence of these 'spacers' does not contribute to autoprocessing, such sequences can be embodied by any combination of any amino acids, as is set forth by SEQ ID NO:31.

With respect to specific sequences, the instant application identifies BIL4_cloth as one specific example of an autoprocessing segment that includes SEQ ID NO:31 (see section [0103] of the published application). The specific amino acid sequence of BIL4_cloth is provided by Table 1 as represented by coordinates 311-345 of contig 23020817. In addition, sections [0188]-[0193] describe construction of the specific sequences used in the experiments. The sequence region containing the autoprocessing segment which was PCR amplified using the primer sequences listed in these sections is described in Table 1 of the instant application.

The Examiner has also rejected claims 1 and 5-18 under 35 USC 112, First Paragraph, as failing to comply with the enablement requirement.

The Examiner states that SEQ ID NO:31 encompasses an extremely large number of different species because of the variability allowed in the sequence.

The Examiner also states that it is well known that amino acid substitutions anywhere in a protein including in regions not required for activity can affect protein structure and function. The Examiner concludes that therefore, undue experimentation would be required of the skilled artisan to make and use the instant invention as claimed.

As is stated above, SEQ ID NO:31 is a specific example of an autoprocessing sequence which includes specific sequence motifs separated by highly variable regions of a specific length.

Variability in 'spacer' regions characterizes autoprocessing domains. For example, the Inteins and Hog Hint domains of the Hint domain family are 130-160 amino acids long and share 4-6 conserved sequence motifs linked via variable intervening (spacer) regions. Although BIL domains belong to the hint domain family, they are distinguished from Inteins and Hog Hints in that they include highly variable 'spacers' since the BIL domains are integrated within non-conserved, hyper-variable proteins (see, http://bioinfo.weizmann.ac.il/~pietro/Hints/). In addition, and as was noted in the previous response, BIL domains are also unique in that they do not require additional flanking sequences for autoprocessing.

Thus, the functional structure of BIL domains includes conserved sequence motifs separated by hypervariable intervening sequences (as exemplified by SEQ ID NO:31).

Studies published following the priority date of the instant application have further substantiated this structure-function relationship of BIL domains by showing that the autoprocessing function of such domains does not depend on the sequence identity of the 'spacer' regions.

For example, Amitai et al. ["Distribution and function of new bacterial intein-like protein domains"Molecular Microbiology 47:61-73 (2003)] and Dassa et al. ["Protein splicing and auto-cleavage of bacterial intein-like domains lacking a C'-flanking nucleophilic residue" J Biological Chemistry 279:32001-32007 (2004)] provide theoretical and computational evidence for the ability of BIL domains to auto process with different 'spacers'. Amitai et al. provide computational evidence by showing the presence of diverse BIL spacers in BIL domains having similar or identical motifs while Dassa et al. constructed a detailed chemical model which illustrates how BIL domains are capable of autoprocessing with diverse spacer sequences.

In re Application of: Pietrokovski et al
Serial No.: 10/534,544
Filed: May 10, 2005
Office Action Mailing Date: Dec 18, 2008

Examiner: Ogunbiyi
Group Art Unit: 1645
Attorney Docket: 29489

35 U.S.C. § 112, Second Paragraph Rejections

The Examiner has rejected claim 10 under 35 USC 112, Second Paragraph as being indefinite.

The Examiner states that claim 10 depends from claim 5 and recites the limitation "wherein said segment of the polypeptide adjacent to said amino terminal end of said autoprocessing segment". The Examiner states that there is insufficient antecedent basis for this limitation.

Claim 10 has now been amended to clearly identify the polypeptide and polypeptide segments referred to.

In view of the above amendments and remarks it is respectfully submitted that claims 1, 5-18 are now in condition for allowance. A prompt notice of allowance is respectfully and earnestly solicited.

Respectfully submitted,

Martin D. Moynihan
Registration No. 40,338

Date: October 13, 2009

Enclosures:

- Petition to Revive an Unintentionally Abandoned Application
- Request for Continued Examination (RCE)
- Two References (Amitai et al. and Dassa et al.)

THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 279, No. 31, Issue of July 30, pp. 32001–32007, 2004
Printed in U.S.A.

Protein Splicing and Auto-cleavage of Bacterial Intein-like Domains Lacking a C′-flanking Nucleophilic Residue*[S]

Received for publication, April 26, 2004, and in revised form, May 17, 2004
Published, JBC Papers in Press, May 18, 2004, DOI 10.1074/jbc.M404562200

Bareket Dassa, Haim Haviv, Gil Amitai, and Shmuel Pietrokovski‡

From the Department of Molecular Genetics, the Weizmann Institute of Science, Rehovot, Israel 76100

Bacterial intein-like (BIL) domains are newly identified homologs of intein protein-splicing domains. The two known types of BIL domains together with inteins and hedgehog (Hog) auto-processing domains form the Hog/intein (HINT) superfamily. BIL domains are distinct from inteins and Hogs in sequence, phylogenetic distribution, and host protein type, but little is known about their biochemical activity. Here we experimentally study the auto-processing activity of four BIL domains. An A-type BIL domain from *Clostridium thermocellum* showed both protein-splicing and auto-cleavage activities. The splicing is notable, because this domain has a native Ala C′-flanking residue rather than a nucleophilic residue, which is absolutely necessary for intein protein splicing. B-type BIL domains from *Rhodobacter sphaeroides* and *Rhodobacter capsulatus* cleaved their N′ or C′ ends. We propose an alternative protein-splicing mechanism for the A-type BIL domains. After an initial N-S acyl shift, creating a thioester bond at the N′ end of the domain, the C′ end of the domain is cleaved by Asn cyclization. The resulting amino end of the C′-flank attacks the thioester bond next at the N′ end of the domain. This aminolysis step splices the two flanks of the domain. The B-type BIL domain cleavage activity is explained in the context of the canonical intein protein-splicing mechanism. Our results suggest that the different HINT domains have related biochemical activities of proteolytic cleavages, ligation and splicing. Yet the predominant reactions diverged in each HINT type according to their specific biological roles. We suggest that the BIL domain cleavage and splicing reactions are mechanisms for post-translationally generating protein variability, particularly in extracellular bacterial proteins.

Bacterial intein-like (BIL)[1] domains are newly identified protein homologs of intein protein-splicing domains (1). The two known types of BIL domains together with inteins and hedgehog-like (Hog) auto-processing domains form the HINT (Hog/Intein) domain superfamily (2). Inteins and Hogs have related auto-catalytic protein-processing activities. Hog domains rearrange their N′-peptide bond into a thioester bond. This thioester is cleaved by a nucleophilic attack of a cholesterol molecule bound by a downstream domain (3, 4). A similar nucleophilic attack occurs during the protein splicing of inteins out of their protein hosts. The rearranged ester/thioester bond at the intein N′ end is attacked by the nucleophilic side chain of the intein C′-flanking residue followed by additional splicing reactions (5). Intein protein splicing thus depends on an invariable Cys, Ser, or Thr nucleophilic C′-flanking residue (+1) for the trans-esterification and acyl rearrangement steps (2, 6).

BIL domains are distinct from inteins and Hogs in sequence, phylogenetic distribution, and host protein type (1). Each of the two BIL types has characteristic and unique sequence features that cluster them separately from other HINT types. Although inteins are integrated in highly conserved sites of essential proteins and Hogs are present in hedgehog and related nematode proteins, BIL domains are integrated in variable regions of non-conserved diverse bacterial proteins, some of which have extracellular motifs. This leads to the hypothesis that BIL domains may have biological roles different from those of other HINT domains (1). Yet little is known regarding the biochemical activity of each BIL type.

We previously described (1) the catalytic activity of an A-type and a B-type BIL domain. The A-type BIL domain was shown to have protein-splicing and C′-cleavage activities. However, this domain was naturally flanked by a Thr +1 residue, which is typical of inteins but not of A-type BIL domains. Only 15% of known A-type BIL domains is followed by Ser or Thr, and none is followed by Cys residues. An A-type BIL domain with +1 Tyr residue was shown recently by Southworth *et al.* (7) to have N′-terminal cleavage but no protein-splicing activity. The B-type BIL domain was examined previously by us only in a cell-free system. It was shown to be active with preliminary evidence for cleavage and protein splicing. Peptide splicing outside the context of intein-like domains also was shown recently to occur in the proteasome, generating variant peptides to be displayed on major histocompatibility complex class I proteins (8).

Here we examine in detail the auto-cleavage and splicing activity of four BIL domains: one A-type BIL domain with a native non-nucleophilic C′-flanking residue (Ala +1) and three different B-type BIL domains. We also show that BIL domains are present in more major groups of bacteria and in proteins likely to be secreted. The probable functions and chemical reaction mechanisms of BIL domains and their relation to inteins are discussed.

EXPERIMENTAL PROCEDURES

Bacterial Strains and DNA Primers—Rhodobacter sphaeroides 2.4.1 (Rsp) genome was a kind gift from Dr. Steven L. Porter (University of Oxford). *Rhodobacter capsulatus* (Rca) MD1 genome was a kind gift from Dr. Fevzi Daldal (University of Pennsylvania), and *Clostridium thermocellum* (Cth) genome was a kind gift from Dr. Ying Tsai (University of Rochester). The following BIL domains were cloned: BIL4-Cth

* The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[S] The on-line version of this article (available at http://www.jbc.org) contains supplemental Figs. S1–S3 and Tables S-I and S-II.

‡ To whom correspondence should be addressed. Tel.: 972-8-9342747; Fax: 972-8-9344108; E-mail: shmuel.pietrokovski@weizmann.ac.il.

[1] The abbreviations used are: BIL, bacterial intein-like; B, BIL; Rsp, *Rhodobacter sphaeroides*; Rca, *Rhodobacter capsulatus*; Cth, *C. thermocellum*; HINT, Hog/intein; MS, mass spectrometry; MALDI, matrix-assisted laser desorption/ionization; M, maltose-binding protein; C, Chitin-binding domain.

TABLE I

Primer name	Primer sequence	Restriction site	Flank amino acids[a]
BIL1-Rsp			
5p-Nrsp-bil1	GAATTCATGGCTGACCAAATCCAGATCGG	EcoRI	+14
3p-Nrsp-bil1	TCTAGAGCGGACGAGGACCCTTTCCGGT	XbaI	+52
BIL2-Rsp+flanks			
5p-rsph-bil2	GAATTCGGTGATTCATCCTTGGGGCGA	EcoRI	+32
3p-rsph-bil2	TCTAGAAAACACGGCAAGGGCGAGCGG	XbaI	+9
BIL2-Rsp-no flanks			
5p-rsp2-bil-only+1	GAATTCCTCTCCCTGACGGCCGGGACG	EcoRI	+1
3p-rsp2-only+1	TCTAGAGGGCCGGGTCACGGGATGGAG	XbaI	+1
BIL4-Cth			
5p1.BIL4 Cth	AAAAGGATCCTGCTTTGTTGCAGGCACGATG	BamHI	
3p408.BIL4 Cth	AAAATCTAGATGCATTATGCACCAATACTTCAT	XbaI	+1
1522-Rca			
5pBIL 1522−108bp	GGATCCAACTACGATCCGACGAACCC	BamHI	+36
3pBIL 1522+105bp	TCTAGAACCATAGCCCTCAAGGCCGTC	XbaI	+35

[a] Number of residues flanking the BIL domain.

(NCBI gi code 23020817); BIL1-Rsp (NCBI gi code 22959584); BIL2-Rsp (NCBI gi code 22959191); and 1522-Rca (1). The BIL domains were amplified by PCR using the primers in Table I and cloned between two protein tags in a plasmid termed pC2C (as described by Amitai *et al.* (1)). This plasmid is a modification of the pMALC2 vector (New England BioLabs, Beverly, MA) containing the *malE* gene for maltose-binding protein (M) from *Escherichia coli* and a downstream *cbd* gene coding for chitin-binding domain (C) from *Bacillus circulans*).

Functional Assay of Protein-splicing and Cleavage Activity—The coding sequence of different BIL domains (B) was cloned in-frame between two protein tags, the maltose-binding protein (M) upstream and the chitin-binding domain (C) downstream. The chimeric protein, M-B-C, was overexpressed and extracted in *E. coli* bacteria as described previously (1). Protein extraction buffer contained 20 mM Tris, pH 7.4, 200 mM NaCl, 1 mM EDTA, and 1 mM sodium azide.

Purification of Tagged Protein Products—Soluble protein products containing either a C or a M tag were purified on affinity columns using chitin (New England Biolabs) or amylose (New England Biolabs) beads, respectively. Lysed cell supernatant in extraction buffer was applied to beads for 1 h at 4 °C with shaking. Elution of proteins from chitin beads was done by mixing the beads with SDS-PAGE sample loading buffer and boiling for 2–3 min. Extraction buffer with 10 mM maltose was used to elute proteins from amylose beads.

Heat Purification of BIL4-Cth Domain—The supernatant of *E. coli* cell lysate overexpressing the BIL4-Cth construct was heated in extraction buffer to 37–80 °C for 20 min. Soluble proteins were separated from the denatured ones by centrifugation at 13,000 rpm for 3 min and applied on an SDS gel.

In Vitro Protein Transcription/Translation—*In vitro* transcription/translation was carried out using *E. coli* S30 extract for circular DNA system (Promega, Madison WI) as described by Amitai *et al.* (1).

Western Blot Analysis—Western blot analysis was used to identify protein products containing either the M or C tag and to identify the GroEL and DnaK protein chaperones. To identify the M tag, monoclonal mouse antibodies directed at maltose-binding protein (Novus Biologicals, Littleton, CO) were used in a 1:800 ratio. To identify the C tag, polyclonal rabbit antibodies directed at CBD (New England Biolabs) were used in a ratio of 1:5000. Antibodies for GroEL were a kind gift from Prof. Amnon Horovitz (rabbit antibodies, used in a ratio of 1:1000), and DnaK mouse antibodies (Stressgen) were used in a ratio of 1:1000. The secondary antibodies used were horseradish peroxidase-conjugated goat anti-mouse IgG or goat anti-rabbit IgG (Jackson ImmunoResearch Laboratories, West Grove, PA) in a ratio of 1:10000. Chemiluminescence detection was held using SuperSignal (Pierce) according to the manufacturer's protocol.

Mass Spectrometry (MS) Methods—Intact molecular weight measurements and peptide mass mapping by matrix-assisted laser desorption/ionization (MALDI) MS were performed at the Weizmann Institute Biological Mass Spectrometry unit and at the Smolar Center for proteins (Technion, Israel). Electroelution from gel followed by in-gel digestion with trypsin, chymotrypsin, or V8 proteases was performed and analyzed as described previously (37).

N-terminal Amino Acid Sequencing—Proteins were electrophoresed by SDS-PAGE, and selected bands were prepared as described by Amitai *et al.* (9) and subjected to Edman degradation at the Weizmann Institute Biological Mass Spectrometry Unit.

Computational Sequence Analysis—Sequence searches used the BLAST programs (10) and the BLIMPS program for block-to-sequence searches (11). Block multiple sequence alignments and phylogenetic analysis were conducted as described by Amitai *et al.* (1). Protein motifs were detected using the InterProScan tool (www.ebi.ac.uk/interpro/scan.html).

RESULTS

To characterize the proteolytic activity of new A- and B-type BIL domains, each BIL domain (B) was cloned in-frame between two protein tags, maltose-binding protein (M) upstream and chitin-binding domain (C) downstream. Protein products of each chimeric gene (M-B-C) were examined *in vivo* and *in vitro* by various methods. To characterize the BIL domain activity in its native protein context, some of the domains were cloned with their full or partial native flanks, whereas others were cloned only with single residue flanks.

Protein Splicing and Cleavage of an A-type BIL Domain with Ala +1 Residue—BIL4-Cth is one of the 23 A-type BIL domains we identified in the thermophilic bacterium Cth (1). It is typical of most A-type BIL domains to have all of the intein protein-splicing active site residues with the exception of the C′-flanking nucleophile (supplemental Fig. S3). Instead of Cys, Ser, or Thr invariably present in inteins, BIL4-Cth is followed by an Ala +1 residue. This is the residue present in 18% A-type BIL domains (fraction calculated as weighted average of putative active domains).

The BIL4-Cth M-B-C precursor was overexpressed *in vivo* as a double-tagged protein, and its products were detected and analyzed. Putative protein-splicing products, the excised BIL domain and the ligated M-C flanks, and the M-B- and M-cleavage products were detected. These products were identified by Western blotting of total cell lysates and affinity-purified proteins separated on SDS-PAGE (Fig. 1*A*). Relative quantities of products were calculated according to measurements taken from Coomassie Blue-stained SDS gels of amylose-purified proteins and total lysates (supplemental Fig. S1). Only trace amounts of the M-B-C precursor were detected under all of the separation procedures, indicating an efficient processing. Spliced product M-C comprised 20–25% of the final products, whereas C′-cleavage product M-B comprised ~5% of the final products. M and B proteins comprised most of the final products, indicating that they were generated by a combination of N′- and C′-cleavages. The final amount of B protein was much larger than the amount of the M-C-splicing product. This finding implies that both protein splicing and cleavage at its N′ and C′ ends released the B protein. The C product was not identified in the gels, perhaps because of cellular degradation.

To characterize the putative splicing product using MALDI

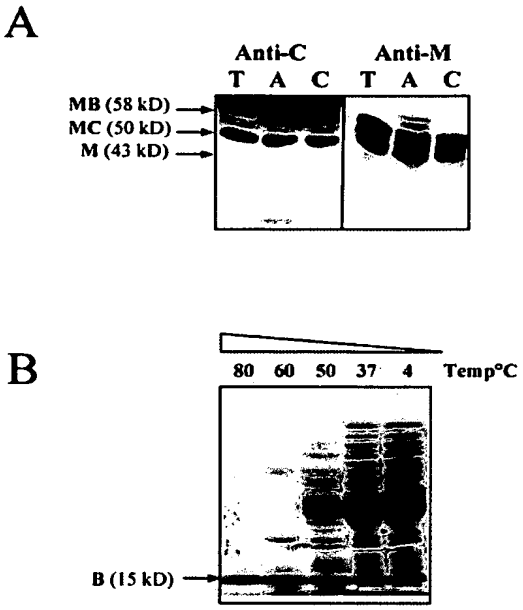

FIG. 1. **Protein-splicing activity of BIL4-Cth.** *A*, Western blotting of *E. coli* overexpression of the M-B-C construct containing A-type BIL4-Cth domain, using anti-maltose-binding protein (*Anti-M*) or anti-CBD (*Anti-C*) antibodies. *T*, total cell lysate; *A*, amylose affinity column eluant; *C*, chitin affinity column eluant. *B*, purification of BIL domain by heat. Samples of total cell lysate were heated to different temperatures. Soluble proteins were isolated and separated on SDS-PAGE.

MS, the M-C band was extracted from the gel and digested with proteolytic enzymes. The presence of the M and C domains was verified using MS/MS analyses. Furthermore, two peptide masses corresponding to splicing-junction peptides were detected from the chymotrypsin digestion of the M-C band. One mass corresponded to a fully cleaved (N'-GSASRVDCG-GLTGL-C') peptide, and another mass corresponded to its mis-cleaved form (N'-GSASRVDCGGLTGLNSGLTTNPGVSAW-C') with high mass accuracies (Table II). The ligated splicing junction is between the second and third residues (Ser-Ala) with the Ser being coded by a linker joining the M tag to the BIL domain and the Ala being the native residue downstream of the BIL domain (Ala +1).

Spliced BIL domain was purified, and its identity was verified by MS. We were able to purify the BIL domain by heat treatment, probably because it originated from a thermophilic bacterium. Incubation of total cell lysate at 80 °C left only the putative BIL domain in the soluble fraction (Fig. 1*B*). Intact mass MS analysis of this 15-kDa band identified the expected mass of the BIL domain, and its sequence was verified by MS/MS analysis (see Table IV and data not shown). The exact C' end of the BIL domain was identified by MS analysis as Asn as expected (Table III).

A putative C'-cleavage product, M-B, was affinity-purified and identified by anti-M antibodies (Fig. 1*A*). Its intact mass analysis corresponded to the expected mass of a C'-cleavage product (Table IV). Other masses obtained from this sample corresponded to the M tag and to other smaller masses that could result from a cross-contamination of the M-B band by traces of smaller proteins on the gel.

A protein band corresponding to the M tag was identified by Coomassie Blue staining and by Western blotting using anti-M antibodies (Fig. 1*A* and supplemental Fig. S1). This putative N'-cleavage product was observed in total lysates and in elutions of both chitin and amylose affinity columns.

To examine whether the Tris cell extraction and protein purification buffer promoted cleavage and splicing of the M-B-C precursor, the extraction and purification procedures were repeated using different buffers (Bis-Tris propane, HEPES, sodium phosphate, and borate). Same products and relative amounts were observed with all of these control buffers (data not shown).

In Vivo and in Vitro Cleavage Activities of B-type BIL Domains—B-type BIL domains are more heterogeneous in sequence than A-type domains (1). To characterize their activity, we cloned three different B-type BIL domains into the double-tagged system (described above): the two BIL domains present in *R. sphaeroides* termed BIL1-Rsp and BIL2-Rsp and one of the 14 BIL domains present in *R. capsulatus* termed 1522-Rca. The conserved C' sequence motif of B-type BIL domains is distinct from those motifs in other known HINT domains (1). The C' end of the cloned BIL1-Rsp and 1522-Rca is typical of B-type BIL domains, whereas BIL2-Rsp has an atypical C' end (supplemental Fig. S3).

N'-cleavage of B-type BIL1-Rsp—BIL1-Rsp, a B-type BIL domain from *R. sphaeroides*, was cloned between M and C tags with its native N'-14 residue and C'-51 residue flanks and overexpressed in *E. coli* cells. M-B-C precursor M and B-C N'-cleavage products were identified by Coomassie Blue staining and Western blotting of total lysate and affinity-purified protein samples (Fig. 2). To verify the nature of the N'-cleavage product, B-C, the band was micro-sequenced. The resulting sequence (*X*FTPGT) corresponded to the predicted N' end of the BIL domain, which also includes Cys-1, which usually cannot be detected by this method (supplemental Table S-I).

An additional 58-kDa band was co-purified with the M-B-C precursor. Its analysis suggests that the band might include more than a single protein species. Both anti-M and anti-C antibodies reacted with this band. However, the peptide mapping of the band identified peptides from both the M tag and the *E. coli* GroEL chaperone protein. Additionally, no peptides from the B and C domains were identified (data not shown). Intact mass of the band identified a mass of 58.317 kDa corresponding to GroEL and an additional unidentified protein mass of 65.175 kDa (Table IV). As a control, we checked a cross-reaction of anti-C antibodies with purified GroEL protein (supplemental Fig. S2*B*). Anti-C antibodies showed reactivity toward GroEL, probably because of their polyclonal nature.

GroEL chaperone was detected in protein samples purified the following affinity columns: on amylose; chitin; and amylose followed by chitin. This indicates a tight and specific binding of GroEL with the precursor and/or protein products. The association of GroEL with unfolded proteins is reversible to some extent upon incubation with ATP-Mg-K (12). Such incubation of washed protein samples bound on chitin reduced but did not eliminate the amount of GroEL eluted from chitin (supplemental Fig. S2*B*).

C'-cleavage of B-type BIL2-Rsp in Vivo, in Vitro, and in Cell-free Systems—BIL2-Rsp was cloned between M and C tags with one native flanking residue at either end (N'-Leu and C'-Pro) and overexpressed in *E. coli* and in a cell-free system. In both systems, the main product was the M-B-C precursor with small amounts of M-B- and M-cleavage products (Fig. 3*A*). An additional band of ~70 kDa appeared above the precursor band when expressed *in vivo*. This band was identified as the *E. coli* DnaK chaperone protein. It was not detected in the overexpressed control protein, M-C. Identity of the above products was verified by Western blotting, N-terminal sequencing of the M-B-C band, MALDI-MS peptide mapping of the M-B and DnaK bands, and MALDI MS intact mass analysis of the M-B band (Fig. 3*A*, Table IV, and data not shown). This last

TABLE II
MALDI-TOF results of BIL4-Cth splicing junction chymotryptic peptides

Sequence position	$[M+H]^+$ calculated mass[a]	$[M+H]^+$ observed mass	Mass accuracy	Sequence
	Da		*ppm*	
392–405	1349.65	1349.71	44.45	GSASRVDCGGLTGL
392–418	2634.26	2634.80	205	GSASRVDCGGLTGLNSGLTTNPGVSAW

[a] Mass calculated with carboxyamidomethyl cysteine modification.

TABLE III
Electrospray Ionization TOF results of BIL4-Cth C-terminal tryptic peptides

Sequence position	Calculated mass	Observed mass[a]	Mass accuracy	Sequence
	Da		*ppm*	
118–135	2108.9545	2109.0533	46.8	VDDFHTYHVGDNEVLVHN
113–135	2760.2927	2760.3261	12.1	VYNFKVDDFHTYHVGDNEVLVHN

[a] Observed masses are an average of $[M+2H]^{2+}$, $[M+3H]^{3+}$, and $[M+4H]^{4+}$ masses for the first peptide and of $[M+3H]^{3+}$ and $[M+4H]^{4+}$ masses for the second peptide.

TABLE IV
MALDI-TOF results of BIL domains splicing and cleavage products intact mass

Clone	Probable product	$[M+H]^+$ calculated mass	$[M+H]^+$ observed mass	Mass accuracy
		kDa		%
BIL4-Cth	M-B	58.038	58.552	0.89
	M	43.092	43.983	2.06
	B	14.963	15.050	0.58
BIL2-RSP-only	M-B	56.782	56.071	1.25
	DnaK	69.118	69.255	0.19
	M-B-C	64.141	65.430	2.00
BIL1-RSP	GroEL/MC	57.332/57.355	58.317	1.72
RCA-1522	GroEL	57.332	57.810	0.83

analysis gave a measured mass of 56.071 kDa, slightly smaller than the expected mass of the putative M-B product.

To examine the *in vitro* activity of BIL2-Rsp, the overexpressed M-B-C precursor was isolated by sequential affinity columns (amylose followed by chitin) and was incubated in the extraction buffer in different temperatures for different time periods. Increasing amounts of the M-B product were clearly detected within 1 day at 4 °C (Fig. 3*B*). The presence of the M band may be attributed to the N′-cleavage of the BIL domain; however, the complementary B-C band was not detected. Alternatively, this could have resulted from protein degradation.

Similar results were observed when BIL2-Rsp domain was cloned with its full native flanks (data not shown). However, this clone also underwent cleavage in an Arg-Arg dipeptide present in the N′-flank of the BIL domain as verified by N-terminal sequencing. This cleavage was also observed when the flanks were cloned without the BIL domain (data not shown). Thus, we suggest that this activity is unrelated to the BIL domain and is probably due to an *E. coli* protease (perhaps OmpT) that can cleave the BIL domain flank.

No Activity of B-type Rca-1522 BIL—1522-Rca B-type BIL domain is natively present in a very large *R. capsulatus* protein. The domain is preceded by 1821 residues and is followed by 52 residues. The upstream flank of this BIL domain includes RTX (repeats-in toxin) calcium-binding repeat motifs, characteristic of secreted proteins (13). The BIL domain was cloned with 36 N′-flanking residues and 35 C′-flanking residues in the double tag expression vector. Overexpression of the vector yielded only the M-B-C precursor and *E. coli* GroEL protein as verified by Coomassie Blue staining, Western blotting, N-terminal sequencing, and MS analysis (Table IV and supplemental Fig. S2*A*).

Isolated M-B-C precursor was incubated *in vitro* at 4 or 37 °C in two different environments of pH 7.4 and 8.5. No products of the precursor were detected under any of these conditions.

Species and Protein Host Distribution of BIL Domains—BIL domains were identified originally in species from Gram-negative α, β, and γ *Proteobacteria* and from Gram-positive *Actinobacteria* and the *Bacillus/Clostridium* group (1). Further data base searches now broaden the taxonomic range of BIL domains to major bacterial divisions and lineages (supplemental Table S-II). A-type BIL domains were found in δ *Proteobacteria*, *Cyanobacteria*, *Spirochaetes*, *Planctomycetes*, and *Verrucomicrobia*. B-type BIL domains were found in α *Proteobacteria*, *Rhizobium*, and *Silicibacter* species.

Sequence analyses of over a hundred identified BIL flanks reconfirmed our previous observation of the nature of the BIL domain hosts. BIL domains are present in homologs of known and predicted secreted proteins. This is exemplified by *Streptomyces avermitilis*, *Verrucomicrobium*, and *Gloeobacter* A-type BIL domains that are found downstream of long (400–5400 residues) Rhs core elements. Rhs elements are composite genetic elements, and their cores are believed to be cell-surface ligand-binding proteins (14). The BIL domains are present in the hypervariable core extension region that can be shuffled between the core and downstream open-reading frame regions.

DISCUSSION

In this study, we show that a typical A-type BIL domain is capable of protein splicing without a C′-nucleophilic +1 residue and that B-type BIL domains can cleave their N′ or C′ ends. Both types of domains are not uncommon, appearing in diverse bacterial divisions. These findings reflect the auto-processing nature of intein-like domains. We explain the N′- and C′-cleavage of B-type BIL domains by reactions occurring in the canonical intein protein-splicing mechanism and propose an alternative pathway for A-type BIL domains splicing. Our results suggest that the biochemical activities of the BIL domains are distinct from inteins, and their native biological function is probably protein modification by splicing and cleavage activity.

Protein-splicing Mechanism without a Nucleophilic +1 Residue—Intein protein-splicing mechanism was largely determined by mutational analysis of a few representative intein domains (2, 6, 15–18). This allowed the delineation of the biochemical reactions of protein splicing and supported splicing as the native activity of inteins. Other evidence for the nature of interin activity are the high efficiency of intein protein-splicing, intein distribution in species and host proteins, and the function of intein genes as selfish genetic elements (19).

Currently, the accepted mechanisms for intein protein-splicing require a Cys, Ser, or Thr +1 residue at the intein immediate C′-flank. This nucleophilic +1 residue is crucial for the trans-esterification step and for the final acyl rearrangement (Fig. 4, *steps 2A* and *4A*). In inteins with N′-Ala-1, the nucleophilic +1 residue directly attacks the peptide bond at the intein N′ end (16). Mutating the intein active site residues, including the +1 nucleophilic residue, abolishes splicing or leads to cleavage of the intein C′, N′ end, or both (15, 20).

In our study, the major products of BIL4-Cth expression were N′- and C′-cleavages, whereas protein splicing was ap-

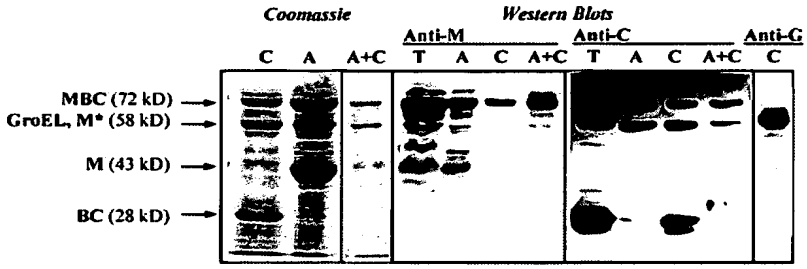

FIG. 2. **N'-cleavage activity of BIL1-Rsp.** Protein products from *E. coli* overexpression of M-B-C construct with B-type BIL1-Rsp were eluted from amylose (*A*), chitin (*C*), or both (*A*+*C*) affinity columns or analyzed in total cell lysate (*T*). Proteins were separated on SDS-PAGE and either stained with Coomassie Blue or detected by anti-M, anti-C, or anti-GroEL (*Anti-G*) antibodies. See "Results" for discussion of GroEL cross-detection by anti-C antibodies.

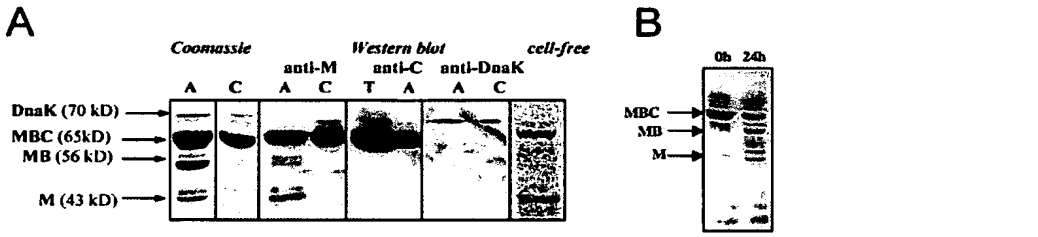

FIG. 3. **C'-cleavage activity of BIL2-Rsp.** *A*, *left*, protein products from *E. coli* overexpression of M-B-C construct with B-type BIL2-Rsp were eluted from either amylose (*A*) or chitin (*C*) affinity columns or analyzed in total cell lysate (*T*). Proteins were separated on SDS-PAGE and either stained with Coomassie Blue or detected by anti-M, anti-C, or anti DnaK antibodies. *A*, *right*, proteins translated *in vitro* in a cell-free system were labeled with [^{35}S]Met. *B*, *in vitro* incubation of a purified precursor at 4 °C.

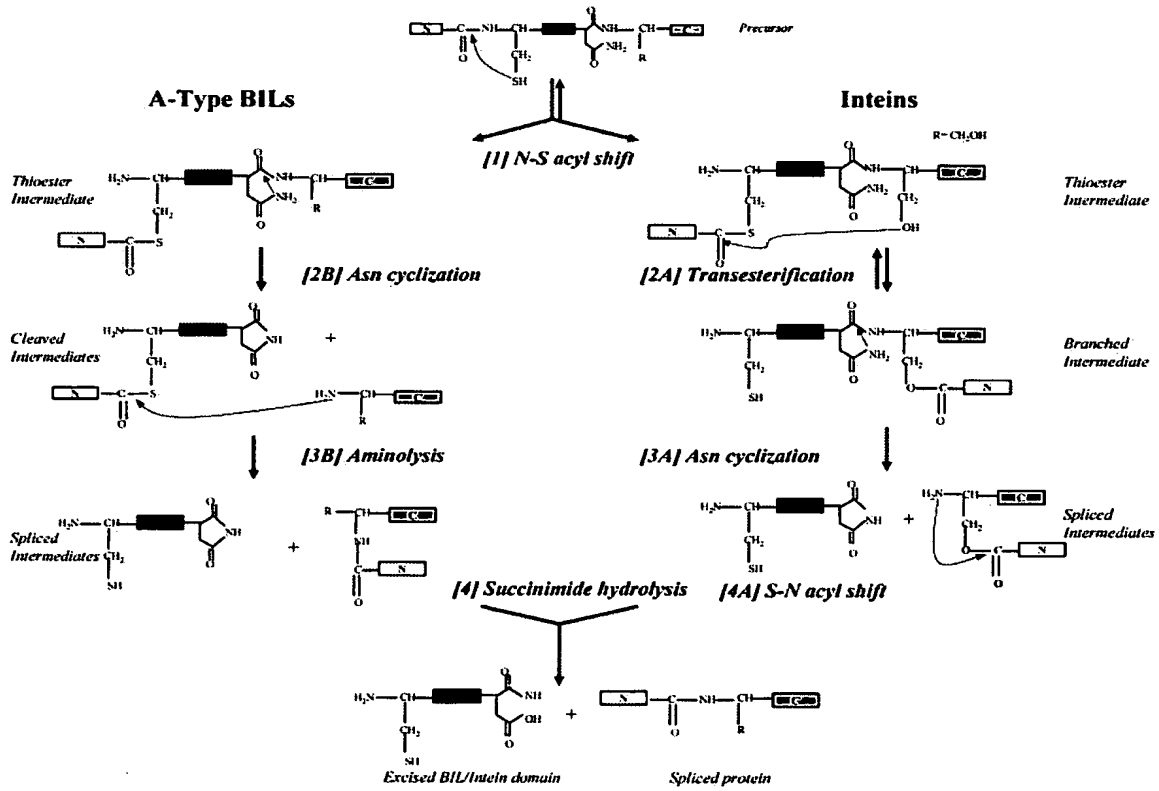

FIG. 4. **Canonical and proposed protein-splicing mechanisms for inteins and A-type BIL domains.** The intein/BIL domain is marked as a *black rectangle* flanked by an N-terminal flank (*N*) and a C-terminal flank (*C*). *Right*, canonical intein protein-splicing mechanism. *Left*, proposed protein-splicing mechanism of A-type BIL domains lacking a C'-nucleophilic residue.

proximately a quarter to a fifth of the A-type BIL domain activity with almost complete processing of the precursor. Most probably, the initial cleavage activity was at the C' end, producing the M-B and C products, followed by additional N'-cleavage of the M-B product, producing the M and B products. This is supported by the relative amounts of the final products and the absence of the B-C product.

Our results show protein splicing of an A-type BIL with conserved sequence features closely related to inteins including all of the active site residues apart from the +1 residue. Hence,

we propose a modified protein-splicing mechanism for A-type BIL domains. The mechanism is similar to the canonical protein-splicing mechanism of inteins, only differing in the nature of the nucleophilic attack on the thioester bond in the N' end at the BIL domain.

Our suggestion includes the following steps of protein splicing in A-type BIL domains (Fig. 4). (i) A thioester is formed at the N' end of the domain by the N-S acyl shift (Fig. 4, *step 1*) by attack of the thiol group of the conserved Cys-1 residue on the carbonyl group of the peptide bond N-terminal to Cys-1. This reaction is the same as the first step of canonical intein protein splicing (15, 18, 20). (ii) Concomitantly, the conserved Asn residue at the C' of the domain undergoes cyclization into an aminosuccinimide ring, cleaving the peptide bond at the domain C' end (Fig. 4, *step 2B*). This step generates two intermediate products: the N'-flank covalently connected to the BIL domain by a thioester bond and the detached C'-flank. This reaction also occurs in intein protein splicing but only after ligation of the two intein flanks (Fig. 4, *step 3A*) (5, 21). In inteins, premature Asn cyclization results in C'-cleavage and no splicing (22). Although the timing of Asn cyclization is tightly controlled in inteins, it can still occur when other steps of the splicing are blocked by mutations at the N'- and/or C'-splice junction (17, 23–25). (iii) The free N terminus of the C'-flank performs an aminolysis reaction of the labile thioester bond next at the N' junction of the domain formed in step i. This reaction ligates the two BIL domain flanks with a peptide bond and releases the BIL domain from its N'-flank. This step probably occurs immediately after step ii to prevent the dissociation of the C'-flank from the N'-flank and BIL domain. (iv) Finally, the BIL domain C'-aminosuccinimide ring hydrolyzes into Asn or iso-Asn, similarly to inteins (Fig. 4, *step 4*) (26).

Aminolysis reaction, involving an attack of the C'-amine on a N'-ester, was proposed previously to occur in intein protein splicing (27, 28). A detailed analysis of representative inteins established the canonical protein-splicing mechanism and ruled out aminolysis as part of the process (15, 20). Considering our experimental results and the various residues in the +1 position of A-type BIL domains, we suggest that these domains protein splice with an aminolysis reaction.

Recently, aminolysis was proposed as part of a peptide-splicing activity of the proteasome that generates the displayed variant antigenic peptides (8). The cleaved peptides within the proteasome are attached transiently from the C' end to Thr residues by ester bonds (21). Vigneron *et al.* (8) suggest that the N' end of another cleaved peptide from the same protein attacks this bond in an aminolysis reaction, ligating the two peptides. Aminolysis also occurs in other biological reactions, including the attachment of myristate to the N' end of proteins by *N*-myristoyltransferase (29).

Why are inteins integrated upstream to Cys, Ser, or Thr residues when, as we show here, protein-splicing can proceed with other residues in this position? Being able to successfully integrate in a wider range of sites seems highly advantageous for selfish genetic elements such as inteins (19, 30). We believe the answer to this question is related to the differences between the mechanisms for protein splicing in inteins and in A-type BIL domains. The intein domain and its flanks remain covalently attached until ligation of the flanks and release of the intein (Fig. 4). In our proposed mechanism for A-type BIL domains, the C'-flank is detached from the BIL and its N'-flank before its ligation to the N'-flank. This may lead to a higher frequency of N'- and C'-cleavage side products. Such partial splicing in inteins will reduce the amount of mature (spliced) host proteins, which are typically conserved, and crucial proteins, and might negatively affect cell survival. Perhaps even more harmful is the possible dominant-negative effect of the cleaved byproducts of intein hosts. In contrast, partial splicing of BIL domains (*i.e.* N'- and/or C'-cleavage) may serve for increasing the protein host variability (1).

Our results, together with previous reports of other atypical intein protein-splicing mechanism (9), show that this activity can proceed by several alternative and partially overlapping biochemical reactions. Thus, the canonical intein protein-splicing mechanism may need to be expanded, or its scope may need to be limited. Aminolysis and perhaps other atypical mechanisms may be the way some inteins and other HINT domains protein-splice.

Cleavage Mechanisms of B-type BIL Domains—The B-type BIL domains were found by us to auto-catalytically cleave their N' or C' ends. This activity is analogous to inteins protein-splicing side reactions and is common in N-terminal rearrangements of auto-processing proteins (2). Both intein and BIL domains have conserved Cys or Ser in position 1 whose thiol or hydroxyl groups are essential for the acyl rearrangement at the N terminus. Thus, the N'-peptide-bond of BIL1-Rsp could be converted into a thioester through the N-S acyl shift, similarly to inteins (Fig. 4, *step 1*). In inteins, this reaction is followed by trans-esterification of the thioester by the side chain of the +1 residue, forming a branch intermediate and leading to splicing product formation. Such products were not obtained in the BIL1-Rsp precursor expression, suggesting that the labile thioester was hydrolyzed by water or by an external nucleophile. We do not exclude the possibility that this cleavage was coupled to ligation of the upstream flank with an external nucleophile, similar to the attachment of cholesterol to Hedge domain upstream to the Hog HINT domain. Such a ligation would modify the M tag and assign it with a higher mass. One of the BIL1-Rsp yet uncharacterized products may correspond to this putative product.

A previously proposed mechanism for C'-cleavage of the *Chy* R1 intein mutant (9) and for *Pab* PolII intein (31) can explain the C'-auto-cleavage of BIL2-Rsp. According to this finding, an attack of the BIL domain Ser-1 hydroxyl group on a peptide bond carbonyl at the C' region of the domain would form an ester bond through the N-O acyl shift, which in turn can be hydrolyzed, detaching the BIL domain from its C'-flank (9). This proposed mechanism is independent of a C'-nucleophilic residue. Assuming that BIL domains have the HINT fold, their N' end is in a position to cleave their C' region.

Our heterologous conditions of protein expression may alter the native activity of BIL domains. Overexpression in *E. coli* cells and changes in the domain context (BIL domain flanks), as well as *in vitro* conditions such as redox environment or temperature, may alter the protein *in vivo* fold and function. Nevertheless, in light of extensive experiments in other proteins and HINT domains, we assume that the BIL domain activity we observed is related to their native one. Improper folding of flanked B-type BIL domains may have triggered the overexpression of chaperones (DnaK, GroEL) (12). We propose that the chaperones, which were co-purified with B-type BIL but were absent in A-type BIL or the control vector, are not merely byproducts of the heterologous expression system. Chaperones may be involved in BIL domains proper folding, extracellular targeting, or biological activity. Attachment of chaperones to the BIL precursor may also spatially block its splicing activity.

Biological Roles of Different Types of HINT Domains—The HINT superfamily currently includes four separate families: inteins; Hogs; A-type BIL; and B-type BIL domains. All of the families are homologous and share sequence, structure, and biochemical properties (2, 4, 6, 32). Yet each family is distinct

in specific sequence features, protein host context, and biological roles. Members of each family can be diverse in sequence and are still found occasionally in new protein and phylogenetic contexts. It is likely that other HINT families will be discovered and characterized. Thus, identifying the family of a HINT domain can be an additional challenge to recognizing the domain as a HINT type.

Sequence motifs and structure folds characterizing the HINT superfamily and those specific to inteins, Hogs, and BIL domains have been described previously (1, 33, 34). Most inteins also include a central homing-endonuclease domain (35) not found in the other known HINT families. Inteins are also integrated in conserved positions of essential proteins. Both these features are a consequence of the selfish element nature of intein genes (19, 30). Hog domains are located upstream to the cholesterol-binding domain and downstream to the Hedge domains and to the Wart and Ground domains of nematodes (36). The role of Hog domains in hedgehog proteins and perhaps also in the nematode proteins is post-translational modification in the maturation process of their host protein.

Less information is available for the two known BIL domains. Nevertheless, the experimental and computational results we show in this work support our initial hypotheses. Most BIL domains are present in variable positions of non-conserved proteins. Many BIL host proteins also include motifs, repeats, and domains that characterize extracellular protein regions. We show here and in the first report of the BIL domains (1) that the biochemical activity of BIL domains includes protein splicing and auto-cleavage of their hosts. We suggest that the biological role of BIL domains is to increase the variability of their hosts, mainly in extracellular protein regions, by cis- and trans-ligation of proteins and other moieties to the hosts.

Acknowledgments—We thank Prof. Meir Wilcheck for supportive suggestions and Prof. Amnon Horovitz for samples of GroEL and its antibodies. We thank the Mass Spectrometry unit of the Weizmann Institute of Science (Rehovot, Israel) and The Smoler Protein Center, Department of Biology (Technion, Israel).

REFERENCES

1. Amitai, G., Belenkiy, O., Dassa, B., Shainskaya, A., and Pietrokovski, S. (2003) *Mol. Microbiol.* **47**, 61–73
2. Paulus, H. (2000) *Annu. Rev. Biochem.* **69**, 447–496
3. Porter, J. A., Ekker, S. C., Park, W. J., von Kessler, D. P., Young, K. E., Chen, C. H., Ma, Y., Woods, A. S., Cotter, R. J., Koonin, E. V., and Beachy, P. A. (1996) *Cell* **86**, 21–34
4. Perler, F. B. (1998) *Cell* **92**, 1–4
5. Xu, M. Q., Comb, D. G., Paulus, H., Noren, C. J., Shao, Y., and Perler, F. B. (1994) *EMBO J.* **13**, 5517–5522
6. Perler, F., Noren, C., and Wang, J. (2000) *Angew. Chem. Int. Ed. Engl.* **39**, 450–466
7. Southworth, M. W., Yin, J., and Perler, F. B. (2004) *Biochem. Soc Trans.* **32**, 250–254
8. Vigneron, N., Stroobant, V., Chapiro, J., Ooms, A., Degiovanni, G., Morel, S., Van Der Bruggen, P., Boon, T., and Van Den Eynde, B. J. (2004) *Science* **304**, 587–590
9. Amitai, G., Dassa, B., and Pietrokovski, S. (2004) *J. Biol. Chem.* **279**, 3121–3131
10. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) *Nucleic Acids Res.* **25**, 3389–3402
11. Henikoff, S., Henikoff, J. G., Alford, W. J., and Pietrokovski, S. (1995) *Gene (Amst.)* **163**, 17–26
12. Bukau, B., and Horwich, A. (1998) *Cell* **92**, 351–366
13. Coote, J. (1992) *FEMS Microbiol. Rev.* **8**, 137–161
14. Yong Dong, W. (1998) *J. Bacteriol.* **180**, 4102–4110
15. Chong, S., Shao, Y., Paulus, H., Benner, J., Perler, F. B., and Xu, M.-Q. (1996) *J. Biol. Chem.* **271**, 22159–22168
16. Southworth, M. W., Benner, J., and Perler, F. B. (2000) *EMBO J.* **19**, 5019–5026
17. Xu, M.-Q., and Perler, F. B. (1996) *EMBO J.* **15**, 5146–5153
18. Romanelli, A., Shekhtman, A., Cowburn, D., and Muir, T. W. (2004) *Proc. Natl. Acad. Sci. U. S. A.* **101**, 6397–6402
19. Pietrokovski, S. (2001) *Trends Genet.* **17**, 465–472
20. Shao, Y., Xu, M. Q., and Paulus, H. (1996) *Biochemistry* **35**, 3810–3815
21. Groll, M., and Huber, R. (2003) *Int. J. Biochem. Cell Biol.* **35**, 606–616
22. Wood, D. W., Wu, W., Belfort, G., Derbyshire, V., and Belfort, M. (1999) *Nat. Biotechnol.* **17**, 889–892
23. Cooper, A. A., Chen, Y. J., Lindorfer, M. A., and Stevens, T. H. (1993) *EMBO J.* **12**, 2575–2583
24. Southworth, M. W., Amaya, K., Evans, T. C., Xu, M. Q., and Perler, F. B. (1999) *BioTechniques* **27**, 110–120
25. Chong, S., Montello, G. E., Zhang, A., Cantor, E. J., Liao, W., Xu, M. Q., and Benner, J. (1998) *Nucleic Acids Res.* **26**, 5109–5115
26. Shao, Y., Xu, M. Q., and Paulus, H. (1995) *Biochemistry* **34**, 10844–10850
27. Clarke, N. (1994) *Proc. Natl. Acad. Sci. U. S. A.* **91**, 11084–11088
28. Perler, F. B., and Adam, E. (2000) *Curr. Opin. Biotechnol.* **11**, 377–383
29. Farazi, T. A., Waksman, G., and Gordon, J. I. (2001) *Biochemistry* **40**, 6335–6343
30. Liu, X. Q. (2000) *Annu. Rev. Genet.* **34**, 61–76
31. Mills, K. V., Manning, J. S., Garcia, A. M., and Wuerdeman, L. A. (2004) *J. Biol. Chem.* **279**, 20685–20691
32. Pietrokovski, S. (1998) *Protein Sci.* **7**, 64–71
33. Hall, T. M., Porter, J. A., Young, K. E., Koonin, E. V., Beachy, P. A., and Leahy, D. J. (1997) *Cell* **91**, 85–97
34. Dalgaard, J. Z., Moser, M. J., Hughey, R., and Mian, I. S. (1997) *J. Comput. Biol.* **4**, 193–214
35. Belfort, M., and Roberts, R. (1997) *Nucleic Acids Res.* **25**, 3379–3388
36. Burglin, T. R. (1996) *Curr. Biol.* **6**, 1047–1050
37. Mehlman, T., Benjamin, M., Merhav, D., Osman, F., Ben-Asouli, Y., Goldshleger, R., Karlish, S., and Shainskaya, A. (2002) *Proceedings of the 50th Conference of American Society for Mass Spectrometry, Orlando, June 2–6, 2002, American Society for Mass Spectrometry*, Santa Fe, NM

SUPPLEMENTAL DATA

Table S-I: N-terminal sequencing results of BIL domains precursor and cleavage products.

Clone	Protein product	Calculated seq	Observed seq
BIL1-RSP	B-C	CFTPGT	XFTPGT
RCA-1522	M-B-C	MKTEEG	MKTEEG
BIL2-RSP(-flanks)	M-B-C	MKTEEG	MN[a]WEEG
BIL2-RSP(+flanks)	B-C	RKGPKM	RKGPKM

[a] A weak signal of Lys was also observed in this position.

Table S-II: Taxonomical distribution of BIL domains

Species and strain	Taxonomic group	No. of BIL domains and type
Rhodobacter capsulatus SB1003	α proteobacteria	14 B [a]
Rhodobacter sphaeroides 2.4.1	α proteobacteria	2 B [a]
Silicibacter pomeroyi DSS-3	α proteobacteria	16 B [a]
Brucella melitensis 16M	α proteobacteria	1 B
Magnetospirillum magnetotacticum MS-1	α proteobacteria	1A, 5B [a]
Methylobacterium extorquens AM1	α proteobacteria	1 B [a]
Rhizobium leguminosarum bv. viciae 3841	α proteobacteria	1 B
Neisseria meningitidis Z2491	β proteobacteria	1 A
Neisseria meningitidis MC58	β proteobacteria	3 A
Neisseria meningitidis FAM18	β proteobacteria	6 A [a]
Neisseria gonorrhoeae FA1090	β proteobacteria	6 A
Chromobacterium violaceum ATCC 12472	β proteobacteria	1 A
Pseudomonas syringae DC3000	γ proteobacteria	1 A [a]
Pseudomonas fluorescens PfO-1	γ proteobacteria	1 A [a]
Pseudomonas fluorescens PfSBW25	γ proteobacteria	1 A [a]
Mannheimia haemolytica PHL213	γ proteobacteria	1 A [a]
Myxococcus xanthus DK1622	δ proteobacteria	3 A [a]
Leptospira interrogans 56601	Spirochaetes	3 A
Streptomyces coelicolor A3(2)	Actinobacteria	1 A
Streptomyces avermitilis MA-468	Actinobacteria	3 A
Thermobifida fusca YX	Actinobacteria	1 A [a]
Clostridium thermocellum ATCC 27405	Clostridia	23 A [a]
Pirellula species 1	Planctomycetes	1 A
Gemmata obscuriglobus UQM 2246	Planctomycetes	2 A [a]
Gloeobacter violaceus PCC 7421	Cyanobacteria	7 A
Verrucomicrobium spinosum DSM 4136	Verrucomicrobia	3 A [a]
Uncultured bacterium 582 clone EBAC080-L028H02	proteobacteria	1 B [a]
Unknown species [b]	unknown	2 B [a]

[a] Genome is not fully sequenced yet, so the number of BIL domains in this strain could possibly be higher.
[b] Sequences from this putative bacteria were DNA contaminants of the *Wolbachia* species *D. melanogaster* genome.

Figure S1

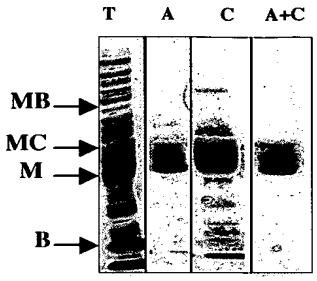

Figure S1. **Protein splicing products of BIL4-Cth**. Coomassie staining of *E. coli* over expression of M-B-C construct containing A-type BIL4-Cth domain. Protein products were eluted from amylose (A), chitin (C) or both (A+C) affinity columns, or analyzed in total cell lysate (T)

Figure S2

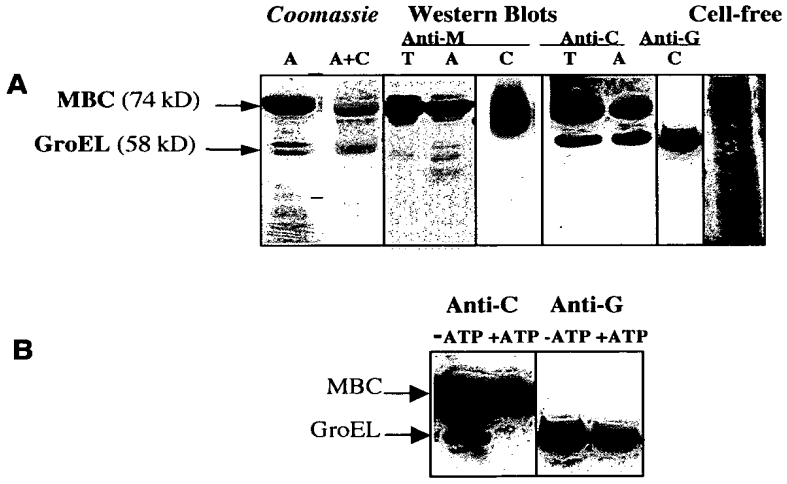

Figure S2. **Experimental analysis of 1522-Rca**. Protein products from in vivo growth were eluted from either amylose (A) or chitin (C) or both (A+C) affinity columns, or analyzed in total cell lysate (T). Proteins were separated on SDS-PAGE and either stained with Coomassie or detected by anti-M, anti-C or anti GroEL (Anti-G) antibodies. Proteins translated *in vitro* in a cell-free system were labeled with S^{35} Met and their autoradiogram is shown.

Figure S3

BIL4-Cth (136aa):
CFVAGTMILT ATGLVAIENI KAGDKVIATN PETFEVAEKT VLETYVRETT
ELLHLTI**GGE VIKTTFDHPF YV**KDVG**FVEA GKLQVGDKLL DSRGNVLVVE**
EKKLEIADKP V**KVYNFKVDD FHTYHV**GD**NE VLVHNA**

BIL1-Rsp (142aa):
CFTPGTLIAT VRGEVAVEAL AAGDRIVTRD NGL**QPLRWIS RRRL**DHATLA
AFPHLK**PVLI EKGSLG**PDLP **DRDMMVSPNH RILV**SRDRTA LHFDAP**EVLV**
AAKHLVGPRG IREVECSG**TT YLHLMFDRHE VVLANGAWTE SF**

BIL2-Rsp (134aa):
SLTAGTPVLT LAGIRPAEGI RPGDRLVARS GAVAVLAAEM TTLPQTE**MVA**
IGASTLAHGQ P**DETLLVPAD QPLLL**RGARA ELLYGQS**PVV LPARRLVDGQ**
LTRLLPMED**V DLVTLTFAAP**
AAIYASELHP VTR

1522-Rca (142aa):
CFTPGTLIA TPKGERLVEEL REGDKILTRD NGI**QEIRWIG RTDL**TRAQLM
ATPHLK**PVLI RAGSLG**NGLP **ERDMLVSPNH RMLV**ANERTA LYFEEH**EVLV**
AAKHLIDNRG VKPVETLG**TS YIHFMFDRHE VVLGNGAWTE SF**

Figure S3. Protein sequence of the analyzed BIL domains. Conserved sequence motifs as described in (1) are marked in bold, including the +1 residue.

Molecular Microbiology (2003) 47(1), 61–73

Distribution and function of new bacterial intein-like protein domains

Gil Amital,[1] Olga Belenkiy,[1] Bareket Dassa,[1] Alla Shainskaya[2] and Shmuel Pietrokovski[1]*
[1]*Molecular Genetics Department and* [2]*Mass Spectrometry Unit, The Weizmann Institute of Science, Rehovot 76100, Israel.*

Summary

Hint protein domains appear in inteins and in the C-terminal region of Hedgehog and Hedgehog-like animal developmental proteins. Intein Hint domains are responsible and sufficient for protein-splicing of their host-protein flanks. In Hedgehog proteins the Hint domain autocatalyses its cleavage from the N-terminal domain of the Hedgehog protein by attaching a cholesterol molecule to it. We identified two new types of Hint domains. Both types have active site sequence features of Hint domains but also possess distinguishing sequence features. The new domains appear in more than 50 different proteins from diverse bacteria, including pathogenic species of humans and plants, such as *Neisseria meningitidis* and *Pseudomonas syringae*. These new domains are termed bacterial intein-like (BIL) domains. Bacterial intein-like domains are present in variable protein regions and are typically flanked by domains that also appear in secreted proteins such as filamentous haemagglutinin and calcium binding RTX repeats. Phylogenetic and genomic analysis of BIL sequences suggests that they were positively selected for in different lineages. We cloned two BIL domains of different types and showed them to be active. One of the domains efficiently cleaved itself from its C-terminal flank and could also protein-splice its two flanks, in *E. coli* and in a cell free system. We discuss several possible biological roles for BIL domains including microevolution and post translational modification for generating protein variability.

Introduction

Hint protein domains appear in two different protein families: inteins and Hogs. In both families the domain performs similar biochemical reactions but in different biological processes. Inteins are selfish genetic elements. They are inserted in frame in various protein coding genes of diverse prokaryotes and few eukaryotes. The whole intein element codes for a protein that is translated with the intein host protein and excises itself from it in a protein-splicing reaction. All inteins have a Hint domain that is responsible and sufficient for the protein-splicing reaction (Paulus, 2000) (Fig. 1). Most inteins also have a homing-endonuclease and DNA-binding domains. These domains mediate the copying of the intein gene into unoccupied intein insertion points (homing). Hog is a C-terminal protein region found in the Hedgehog developmental proteins of vertebrates and insects and in three other nematode protein families (Aspock *et al.*, 1999). The Hint domain is the N-terminal part of Hog regions and in Hedgehog proteins it is followed by a cholesterol-binding domain. In Hedgehog proteins the Hog region mediates its cleavage from the N-terminal domain of the protein (Hedge) by attaching a cholesterol molecule to the Hedge domain (Fig. 1). This activates the Hedge domain that is then secreted from the cell (Porter *et al.*, 1996a, b).

The first biochemical reactions of intein excision from its protein host and Hog detachment from its N-terminal domain are identical (Paulus, 2000): the peptide bond N-terminal to the end of the Hint domain is converted into a thioester (or ester) bond. A *trans*-esterification reaction then attaches the C-terminal flank of the intein, or a cholesterol molecule in Hog proteins, to the N-terminal domain cleaving it from the Hint domain. Hint domains in both families also have the same structure fold and sequence motifs (Dalgaard *et al.*, 1997; Hall *et al.*, 1997; Pietrokovski, 1998). Interestingly, the phylogenetic distributions of inteins and Hog proteins, known so far, do not overlap. Inteins are found in prokaryotes, single cell eukaryotes, plastids and viruses (Pietrokovski, 2001) and Hog proteins are found in multicellular animals (Aspock *et al.*, 1999). The divergence of the Hog protein Hint domains from intein Hint domains is estimated to have occurred at, or prior to, the appearance of metazoa (Pietrokovski, 2001).

Hint domains are necessary for the maturation of the Hedgehog proteins in which they are found, and perhaps this is the role of Hint domains also in other Hog proteins

Accepted 19 September, 2002. *For correspondence. E-mail pietro@bioinfo.weizmann.ac.il; Tel. (+972) 8 934 2747; Fax (+972) 8 934 4108.

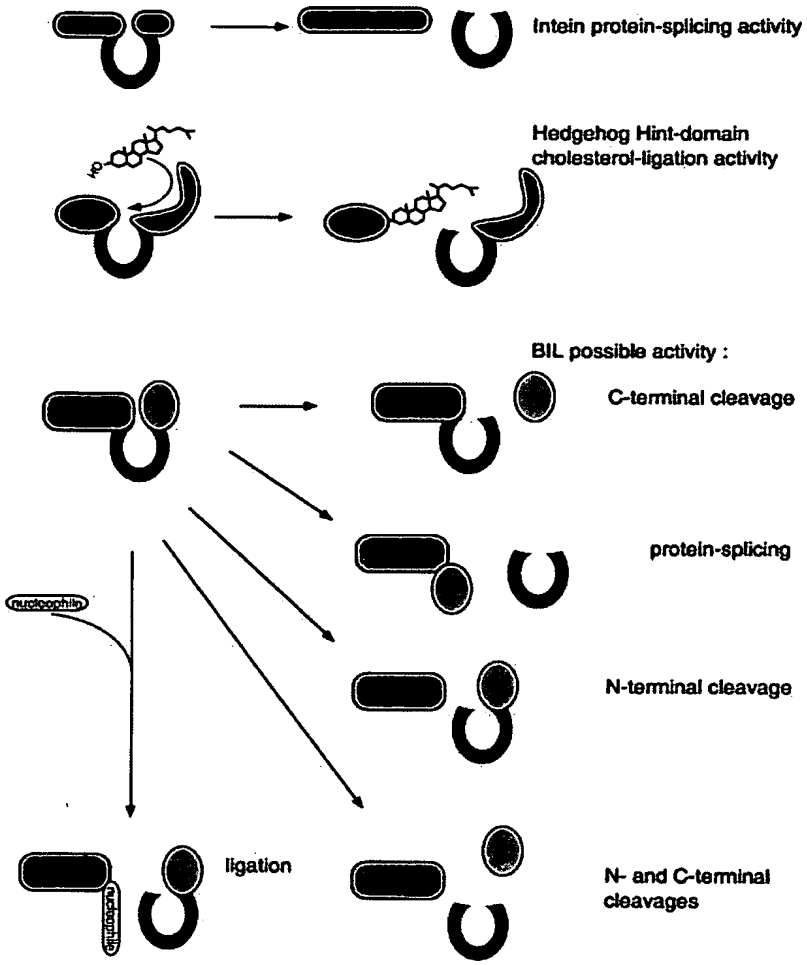

Fig. 1. Known and possible activity of Hint domains.
Top. Known protein splicing of inteins and cholesterol-ligation dependent N-terminal cleavage of hedgehog Hint domains.
Bottom. Possible functions of BIL domains.
Intein, Hedgehog and BIL Hint domains are shown as dark grey horseshoes with their flanks as ovals. The Hedgehog cholesterol binding domain is shown stippled. The proteins N-terminal ends are on their left.

(Porter *et al.*, 1996a). Inteins, the progenitors of Hog proteins Hint domains, are selfish genetic elements based on all current evidence (Pietrokovski, 2001). This raises the issue whether the origin of inteins themselves is from proteins that had a different biological role, and what is the nature of that role. Specifically, protein-splicing could modulate the molecular function of host proteins although no intein is known to do so.

Here we report the identification of various bacterial proteins that include two new types of Hint domains. The phylogenetic and genomic distributions of these domains are analysed. We show that two of these domains, one from each type, are active. One of the domains can protein splice and we suggest that the role of BILs is to process proteins. In at least some species, including pathogenic bacteria, this processing could increase the variability of secreted proteins. This might be a new mechanism for generating protein variability.

Results

New Hint-like protein domains in bacteria

Database searches for protein sequences with Hint domains identified more than 50 such open reading frames (ORFs) in diverse bacterial species, Table 1. These Hint domains are termed BILs for bacterial intein-like domains. Bacterial intein-like domains are 130–155 aa long and have characteristic sequence motifs of the Hint domain (Pietrokovski, 1994; 1998; Hall *et al.*, 1997) (Fig. 2). These domains are distinct from inteins in having additional unique sequence motifs, not being integrated in highly conserved sites of essential proteins and in not including endonuclease domains. They are also distinct from Hog protein Hint domains by: (i) lacking conserved motifs characteristic to those domains; (ii) occurring in bacteria; and (iii) being flanked by protein domains unlike those found in all known Hog proteins. Two types of BILs

Table 1. Distribution of bacterial intein-like (BIL) domains.

Species and strain	Taxonomic group	BIL number and type
Rhodobacter capsulatus SB1003	α proteobacteria	14[a] B
Rhodobacter sphaeroides 2.4.1	α proteobacteria	2[a] B
Brucella melitensis 16 M	α proteobacteria	1 B
Magnetospirillum magnetotacticum MS-1	α proteobacteria	1[a] A 5[a] B
Neisseria meningitidis Z2491	β proteobacteria	1 A
Neisseria meningitidis MC58	β proteobacteria	3 A
Neisseria meningitidis FAM18	β proteobacteria	6 A
Neisseria gonorrhoeae FA1090	β proteobacteria	6 A
Pseudomonas syringae DC3000	γ proteobacteria	1[a] A
Mannheimia haemolytica PHL213	γ proteobacteria	1[a] A
Streptomyces coelicolor A3(2)	Actinobacteria	1 A
Thermobifida fusca YX	Actinobacteria	1[a] A
Clostridium thermocellum ATCC 27405	*Bacillus/Clostridium* group	10[a] A

a. Genome is not fully sequenced yet, so the number of BILs in this strain could possibly be higher.

(termed A and B) are apparent by specific sequence motifs and by features of motifs common to both BILs (Figs 2 and 3 and see also *Supplementary material* Fig. S1).

BILs are found in α, β and γ proteobacteria (Gram-negative bacteria), in actinobacteria (high GC Gram-positive bacteria) and in the *Bacillus/Clostridium* group (low GC Gram-positive bacteria). Both their presence and genomic distribution are variable, even in closely related species and strains. For example, in three complete and almost complete sequenced strains of *Neisseria meningitis* there are one, three or six ORFs with BILs, in two different *Rhodobacter* species there are two and 14 ORFs with BILs, and whereas one strain of *Pseudomonas syringae* has one such ORF, *Pseudomonas aeruginosa* and *Pseudomonas putida* have none. Different BIL types and inteins co-exist in certain species, i.e. *Magnetospirillum magnetotacticum* has both A- and B-type BILs and *Thermobifida fusca* has both BILs and inteins, Table 1.

The variability in number of ORFs with BILs in different species is probably due to gene duplications. Dendograms of BIL domains show that all those derived from *Neisseria* species cluster together and BILs from different species subcluster as well. This implies that all *Neisseria* BILs duplicated from one ancestor and some are paralogues within different species (Fig. 3). This is corroborated by the apparent duplication of some gene loci with BIL ORFs in these species (not shown). Clustering of BILs from the same species is also found in *Clostridium thermocellum* and *Magnetospirillum*.

Properties of BIL-domain proteins

BILs are present in ORFs that can code for a few hundred to a few thousand amino acids. Some of the shorter ORFs might be non-functional genes because they include in frame stop codons and no clear promoter sequences and translation initiation signals were found. In addition, a few ORFs in *C. thermocellum* and different *N. meningitidis* strains are truncated, missing N-terminal parts of the BIL and its N-terminal flank (see *Supplementary material* Fig. S1).

Several BILs are flanked by domains that are present in secreted bacterial proteins. In *P. syringae* and *Mannheimia haemolytica* BILs are found in FhaB-like ORFs, near their C-termini. FhaB is an extremely large *Bordetella* gene, coding for a protein of a few thousand amino acids that is a secreted filamentous haemagglutinin. It functions as an adhesin and is important for *B. pertussis* virulence (Smith *et al.*, 2001). Three of the *Rhodobacter capsulatus* ORFs with BILs include RTX repeats. These calcium binding repeats are found in various secreted bacterial proteins, including many toxins (Coote, 1992). In *N. meningitis* and *N. gonorrhoeae* some BILs are found in MafB proteins. These are part of multiple adhesin family possibly involved in glycolipid adhesion to cells (Paruchuri *et al.*, 1990; Naumann *et al.*, 1999). Three other Neisserial BILs have an HNH nuclease domain in their C-terminal flanks. HNH domains appear in various DNase and endonuclease proteins including secreted toxins (James *et al.*, 1996; Belfort and Roberts, 1997). A domain present in the C-terminal flank of a BIL in the gram-positive bacterium *T. fusca* is also found in a *Salmonella* short conserved ORF (GenBank accession NP_454902) and in the C-terminus of a *N. meningitis* FhaB/Haemolysin protein (gene NMA0688). Both these proteins are from Gram-negative bacteria and are likely to be secreted.

Features of BIL domains

Identification of Hint domain motifs in both BIL types enabled us to locate BIL residues that correspond to the intein protein-splicing active site. Generally, these motifs are conserved and similar in nature to those appearing in

Fig. 2. Conserved motifs of Hint protein domains. Each row shows conserved motifs from one type of Hint protein domain. Motifs are ordered left to right in the N′ to C′ positions along the protein sequences. Similar motifs are vertically aligned with each other. Unique A-type and B-type BIL motifs are underlined with hatched lines. The motifs are shown as sequence logos where the height of amino acids are proportional to their conservation in each position. Positions of the intein protein-splicing active site residues are marked by asterisks. Protein motifs were found and are displayed as previously described (Pietrokovski, 1998). The BIL motif sequences and the distances between consecutive motifs are listed in supplementary Table 1. Intein and hedgehog Hint domains are those described by Aspock (1999) and Pietrokovski (2001). Only intein and hedgehog motifs common to Hint domains are shown.

inteins and Hog proteins Hint domains (Fig. 2). However, B-type BILs are missing the C-terminal motif of inteins. Their final three conserved C-terminal residues are different from those present in inteins. One possible resemblance to the C-termini of inteins is the C-terminal penultimate position of B-type BILs. It is either a cysteine, serine or threonine. This is the conservation found in the position following the C-end of inteins. The SH/OH groups on the side chains of these residues in intein host proteins are crucial for ligation of the intein C- and N-flanks in the protein-splicing reaction (Xu and Perler, 1996). At present, we cannot ascertain what is the exact C-terminal end of B-type BILs, and some conserved positions at that region might belong to the BILs C-terminal flank. In A-type BILs, the two C-terminal residues are invariant histidine-asparagine, identical to the typical C-terminal residues of inteins. However, the following position in A-type BILs, which corresponds to the first residue of inteins C-terminal flanks, is not conserved (Fig. 2).

Almost all A-type BIL domains have apparent functional protein-splicing active sites, and a few also have flanking C-terminal residues found in inteins (serine and threonine). In order to verify protein-splicing activity we cloned and expressed an A-type BIL domain from *P. syringae*. This domain has typical intein residues in all active site residues (Fig. 4). To examine what is the activity of B-type BIL domains we also cloned one such domain from *R. sphaeroides*.

Experimental analysis of A-type and B-type BIL domains

To experimentally test BIL domain activity we cloned an A-type and a B-type BIL domains each between two tag domains. We then examined the size and nature of the resulting proteins in an *in vitro* translation system and after expression in *E. coli*.

A gene encoding a chimeric protein, MBC, composed of the maltose-binding protein (M 43045.8 Da), *P. syringae* (Psy) A-type BIL domain and its downstream threonine (B 16024.2 Da), and chitin binding domain (C 7201.5 Da) was constructed in an expression plasmid.

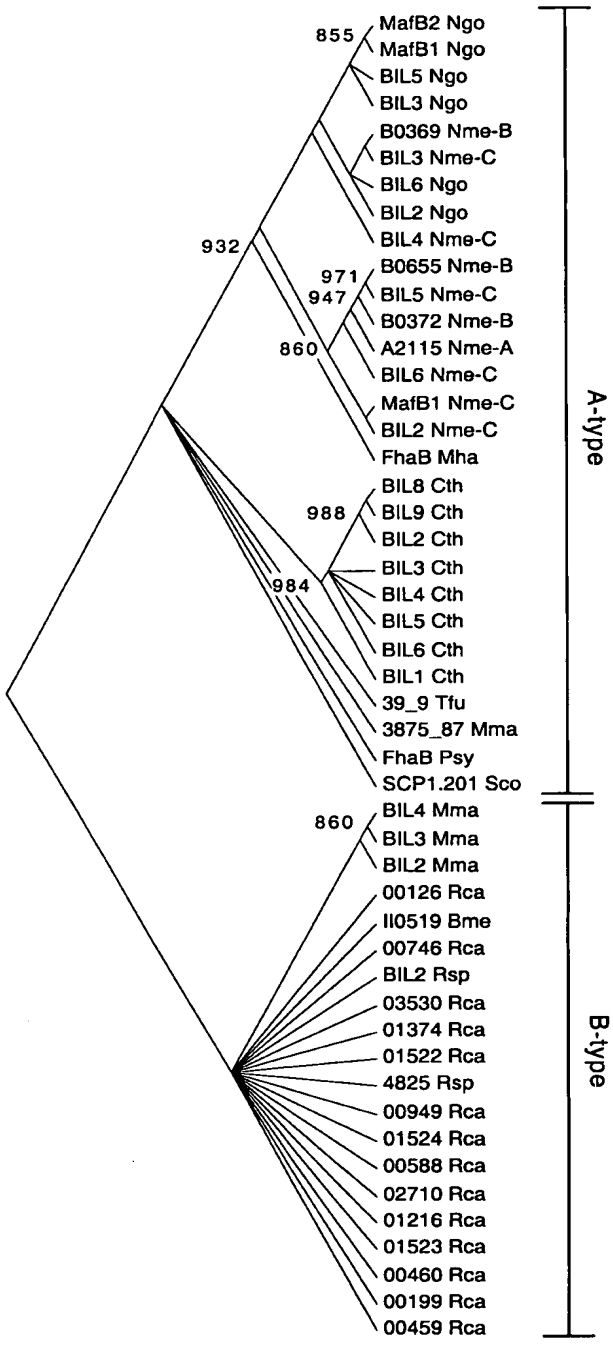

Fig. 3. BIL domains dendogram. Dendogram was computed from a DNA multiple sequence alignment of 49 mostly complete BIL domains (Table 1), aligned across 201 positions, coding for 67 amino acids that could be confidently aligned across the A-type and B-type sequences (Fig. 2). Nodes with bootstrap values below 440/1000 were collapsed. Bootstrap values above 800/1000 are shown. Bootstrap values of the nodes grouping all A-type and B-type BILs are 441 and 519, respectively and the *D. melanogaster* Hedgehog Hint domain (Porter *et al.*, 1996a) was used as an outgroup to root the tree. The dendogram was calculated by the DNADIST program (version 3.5) of the PHYLIP package(Felsenstein, 1989). Similar results were found by the CLUSTALW program (Thompson *et al.*, 1994) and from the protein multiple sequence alignment by the PHYLIP protdist and by the CLUSTALW programs. Species are named as follows: Nme-A *Neisseria meningitidis* strain Z2491, Nme-B *Neisseria meningitidis* strain MC58, Nme-C *Neisseria meningitidis* strain FAM18, Ngo *Neisseria gonorrhoeae* strain FA1090, Mha *Mannheimia haemolytica*, Tfu *Thermobifida fusca*, Mma *Magnetospirillum magnetotacticum*, Psy *Pseudomonas syringae*, Scp *Streptomyces coelicolor* strain A3(2), Cth *Clostridium thermocellum*, Rca *Rhodobacter capsulatus*, Rsp *Rhodobacter sphaeroides*, Bme *Brucella melitensis*.

Protein splicing of the MBC precursor would produce MC and B proteins, C-terminal BIL cleavage would produce MB and C proteins, and N-terminal BIL cleavage would produce M and BC proteins. The plasmid was expressed by an *in vitro* transcription/translation system. Five distinct protein products were identified (Fig. 5A). Three protein bands had weights corresponding to the MBC precursor, MC the protein splicing product and MB the C-terminal cleavage product. The two additional bands had similar weights of 43 and 45 kDa. Control reactions with a template of an MC protein (a plasmid without the BIL insert) yielded two protein bands. One corresponded in weight to the MC protein and the other to the maltose-binding protein (43 kDa). Such a product can be seen in chimeric proteins having the M protein as an N-terminal tag (i.e. NEB instruction manual 'pMAL protein fusion and purification system', Catalogue #E8000S).

The 43 and 45 kDa weight bands identified from the Psy MBC gene are thus considered as a premature transcription or translation stops side product, not related to the BIL domain activity. Appearance of the 45 kDa band, not seen in the control reaction and slightly larger than the expected weight of M, maybe the result of an additional termination point introduced in the BIL domain. Radioactive methionine was used to label the reaction products. Unlike the M and B domains, the C domain has no methionines and therefor its product without M or B domains could not be labelled this way.

The approximate relative amounts of the three products considered specific to complete translation of the BIL containing protein were MBC 15%, MB 57% and MC 28% (Fig. 5A).

The B-type *R. sphaeroides* (Rsp) BIL2 domain (supplementary Table S1) was similarly cloned in between the M and C domains but with its 32 aa and 11 aa N- and C-terminal flanks respectively. Both M and C domains now had additional flanks and the sizes of the three domains of the Rsp MBC protein were: M 46109.48 Da, B 13895.03 Da and C 8061.52 Da. Following *in vitro* translation the products included bands with

CFAAGTMVST PDGERAIDTL KVGDIVWSKP EGGGKPFAAA ILATHIRTDQ PIYRLKLKGK 6046

QENGQAEDES LLVTPGHPFY VPAQHGFVPV IDLKPGDRLQ SLADGASENT SSEVESLELY 6106

LPVGKTYNLT VDVGHTFYVG KLKTWVHNT 6135

Fig. 4. *P. syringae* FhaB BIL domain sequence. Protein sequence of the *Pseudomonas syringae* FhaB BIL domain. Regions corresponding to the six protein-splicing motifs (Pietrokovski, 1998) are underlined with active site residues double underlined. Co-ordinates show the domain position in the FhaB protein. The length of the protein is estimated to be 6274 amino acids long, with the exact position of the N-terminal end uncertain.

sizes corresponding to the MBC precursor (65 kDa), to the M domain (43 kDa), that also appeared in the control reaction, and bands with sizes corresponding to MB (61.5 kDa) and MC (56 kDa) proteins of this construct (Fig. 5B).

To better quantify the Psy BIL reaction products and examine it in an *in vivo* system, the Psy MBC protein was over expressed in *E. coli* and affinity-purified by either of its two protein tags. The purified proteins again included bands corresponding to spliced- (MC) and C-terminal cleaved products (MB) by both predicted mass and Western-blot analysis (Fig. 6A). The main product was again the MB protein, as displayed by comparing its amount to that of the MC protein when both were purified on amylose beads (Fig. 6A, left panel lane 3).

Identities of the MC and MB Psy BIL protein products were confirmed by mass spectrometry analysis (Fig. 6B). The measured mass for the MC protein (50 602.07 Da) is in close agreement to the expected mass for an unmodified protein (50 266.39 Da). The MB protein measured and expected masses are also in close agreement: 59 332.79 and 59070.11 Da respectively. A prominent peak with a mass of 43 303 Da, also observed in MALDI spectra of electroeluted 50 kDa MC fragment is probably a cross-contamination by traces from the lower mass band observed on the gel (Fig. 6A). Such cross-contamination can be observed in gel purified protein bands (A. Shainskaya, unpublished). Reactivity of the lower mass band with antibody against the M-tag (see Fig. 6A), indicates that this ~43 kDa band is a truncated product.

Peptide mass mapping of the MC and MB Psy BIL reaction products by MALDI analysis further validated their assignment. In particular, it identified the splice junction of the MC protein and the C-terminal end of the MB protein. A peptide corresponding to the ligated ends of the M and C protein tags (N'-ISEFGSTSR-C') was identified in the MC protein with an accuracy of 71 p.p.m. Overlap-

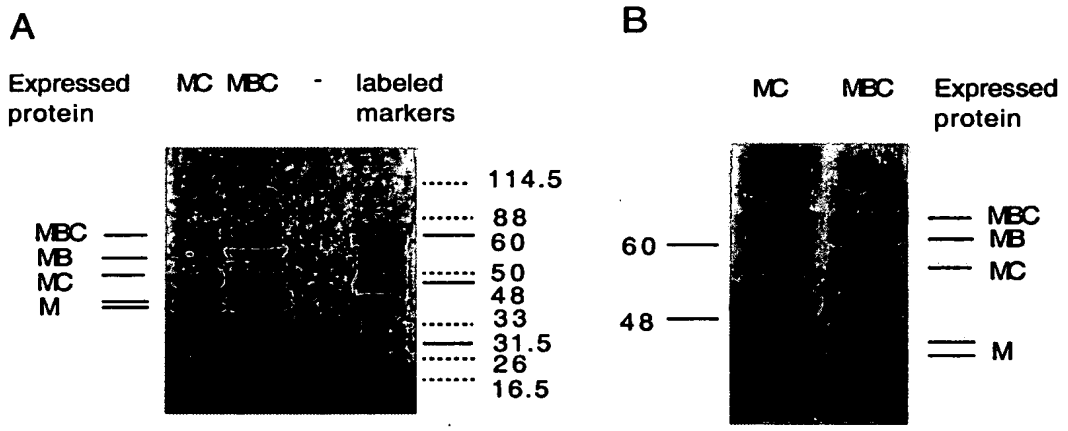

Fig. 5. *In vitro* translation products of a fusion protein including BIL domains.
A. Translation of a chimeric MBC protein with the *P. syringae* FhaB BIL domain.
B. Translation of a chimeric MBC protein with the *R. sphaeroides* BIL2 BIL domain.
PAGE separation of *in vitro* translated and [^{35}S]-methionine labelled pC2C-PsyBIL (A) and pC2C-RspBIL2 (B) plasmids MBC genes. Translation of plasmids pC2C, pBEST*luc* and of no DNA served as controls. Molecular weights, in Kda, were estimated by the products of the pBEST*luc* plasmid and, in A, by unlabelled protein markers run on the same gel, marked with dotted lines. Expressed protein marks our identification of the pC2C-PsyBIL and pC2C-RspBIL2 translation products.

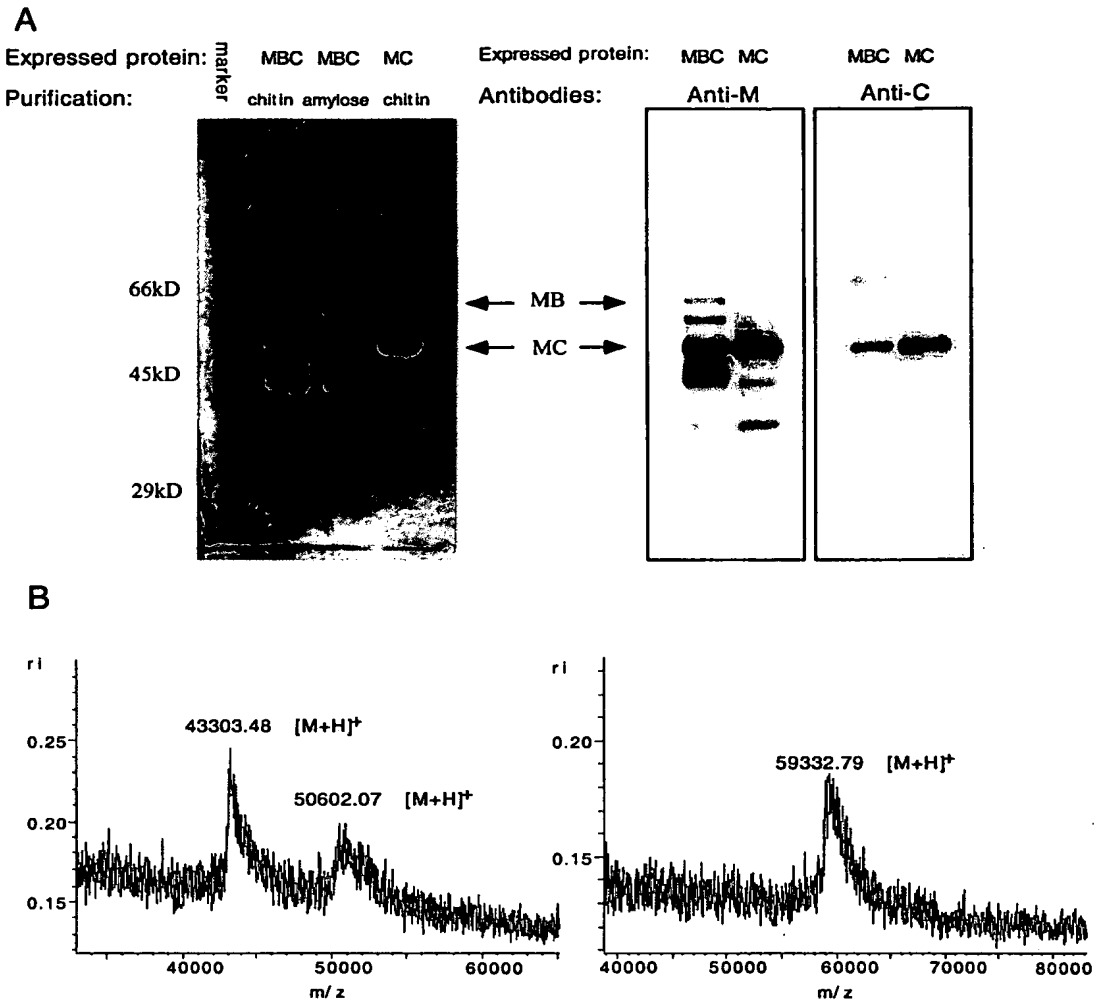

Fig. 6 *In vivo* products of a fusion protein including the *P. syringae* FhaB BIL domain.
A. left panel, PAGE separation of overexpressed pC2C-PsyBIL plasmid MBC protein product. The second and third lanes are proteins purified on chitin and amylose affinity columns respectively. The fourth lane is a control showing overexpression of pC2C plasmid MC protein product. Identity of protein bands from these lanes is shown on the right. Protein markers are shown on the first lane with the masses of protein bands. The top band from the 3rd lane and the central band from the 2nd lane, corresponding to the masses of MB and MC, respectively, were excised from the gel and analysed by MALDI-TOF mass-spectrometry. Right panel, Western-blot analysis of the overexpressed MBC protein after purification on chitin beads. Both chitin-purified samples were run on duplicate lanes and blotted to a single nitrocellulose membrane. The membrane was cut in half and each sample duplicate reacted with either anti-MBP or anti-CBD antibodies. Both antibodies against MBP and CBD tags reacted with the protein band corresponding to the mass of the MC product. Protein bands corresponding to MB and M products, that appear after purification on chitin beads probably result non-specific binding by excess amounts of overexpressed protein.
B. MALDI mass spectra of the protein products MC, left, and MB, right, electroeluted from Coomassie stained SDS-PAGE gels.

ping peptides corresponding to the C-terminal end of the Psy BIL domain (N′-LKTWVHN-C′ and N′-TWVHN-C′) were identified in the MB protein with accuracy of 27 and 100 p.p.m., respectively. Additional peptides from both M and C protein tags have been identified in agreement with the assignment of the MC and MB products from the SDS-PAGE, Western blot and MALDI-TOF analyses (see *Supplementary material* Figs S1 and S2).

Discussion

BILs are new and distinct Hint domains

We have identified in three distant groups of bacteria (proteobacteria, actinobacteria and the *Bacillus/Clostridium* group) a protein domain that is homologous to Hint domains. This new domain appears in non-conserved regions of hypervariable proteins. Thus, these domains

appear distinct from the Hint domains of inteins and Hog-proteins by the species and proteins in which they appear. The BIL (bacterial intein-like) domains are separated by sequence features into A- and B-type domains.

BIL domains are also distinct from inteins by their global sequence features. We examined BLAST sequence searches (Altschul *et al.*, 1997) of BIL domains with BIL domains and with intein sequences. BIL domains were aligned with other BIL domains with higher scores than with intein sequences and their alignments with each other was across their whole, or almost whole, lengths (results not shown). This is also a practical way to distinguish these two related domains from each other.

Sequence to function analysis of BIL domains

The protein-splicing active site residues are present and conserved in BIL domains with one exception. The C-terminal ends of both BIL types each differ from that of inteins. In A-type BIL domains, an absolutely conserved histidine-asparagine motif is present at the C-terminal end, identical to the typical C-terminal positions of inteins. However, the C-terminal flanking position of this motif is not well conserved. Whereas all inteins have only cysteine, serine or threonine in this position, just a few A-type BIL domains have serine or threonine in that position. Other A-type BIL domains have aspartate, glutamate, asparagine, tyrosine or alanine residues in that position, which are not found in any intein (Fig. 2). B-type BILs C-terminal ends have a conserved position of cysteine, serine or threonine. This could correspond to the C-terminal flanking position of inteins but it is not preceded by the histidine-asparagine motif typically found in inteins (Fig. 2).

Conservation of the A-type BIL C-terminal end suggests cyclization of the C-terminal asparagine residue in the same manner as in inteins. We proved that in an A-type BIL the peptide bond between this asparagine and the following threonine is cleaved. In other A-type BILs that do not have C-terminal flanking threonine or serine residues, asparagine cyclization might occur without *trans*-esterification by the flanking residue. It is also possible that *trans*-esterification occurs by mild nucleophilic residues found in this position. In the first case the BIL would be cleaved from its C-terminal flank and in the second protein-splicing will occur.

B-type BIL domains do not have any conserved asparagine or glutamine residue at their C-terminal end. Cleavage of this end could then proceed by a mechanism different from the asparagine and glutamine cyclizations of inteins (Paulus, 2000). Alternatively, B-type BIL domains might not be cleaved at their C-terminal ends, similarly to Hedgehog Hint domains.

Activity of BIL domains

The *P. syringae* BIL domain that we cloned was active both in a cell free system and *in vivo*. In both systems the BIL activity mainly resulted in a single cleavage at the BIL C-terminal end. Lesser amounts of a protein-splicing product were also produced. It is possible that N-terminal cleavage also occurred as M domain product was detected in the two systems. However, in both cases no complementing BC product was identified. If N-terminal cleavage occurred than its BC product was unstable in both systems. We believe this is not very likely and consider the M domain products to be the result of immature transcription and/or translation products or of protein degradation. Our results clearly demonstrate that a BIL domain can protein-splice and cleave its C-terminal end. Both reactions are probably autocatalytic as they readily occurred in a cell free system.

Pseudomonas syringae BIL domain has both sequence features and splicing function of inteins. This indicates the natural molecular function of at least certain BIL domains. Function of other A-type BILs, less similar to inteins, and of B-type BILs might be different, as discussed above. We examined the domain function in a protein context different from its natural setting. Inteins are generally believed to have native activity at these conditions. Further support for the relevancy of our results to the natural BIL activity comes from finding the same activity in both *in vivo* and *in vitro* expression systems. Nevertheless, more experiments in different contexts (e.g. natural flanks and natural species) are needed to verify the 'wild type' molecular function of various BILs.

The *R. sphaeroides* BIL domain we cloned was active in a cell free system. Our preliminary evidence indicates C-terminal cleavage and protein splicing showing this domain is active. Further experiments are needed to better characterize the activity of this domain. Because B-type BIL domains have characteristic conserved features this may indicate most would have some activity.

Activity of both tested BIL domains in the cell free system also supports our claim for post translational modifications versus the alternative for RNA splicing. The later possibility has not been rigorously disproven but is unlikely for several additional reasons. First, no sequence features of any type of intron is found in the genes we studied. The splicing junction and cleavage point of the Psy BIL domain are also exactly those predicted from the sequence similarity of the BIL and intein domains (see *Supplementary material* Figs S1 and S2). Appearance of both splicing and cleavage products is also often found in studies of inteins (Paulus, 2000).

Possible biological functions of BIL domains

Sequence similarity between BILs from the same species

(Fig. 3) and the presence of BIL gene clusters indicate their expansion, and thus positive selection, within some species. We propose that this selection is for the BILs function and that they do not serve as mere static 'spacer' domains. Our demonstration of protein-splicing and cleavage activity of the *P. syringae* BIL domain implies the presence of these activities in other BILs. This stems from the fact that residues forming the protein-splicing active site in inteins are also conserved on most other BILs.

BILs are present in several hypervariable bacterial proteins, such as FhaB adhesins and MafB Neisserial proteins. Their immediate flanks are the most variable portions of the proteins and they themselves are not always present in these proteins, even in closely related strains of the same species. Some, and perhaps all, proteins with BIL domains seem to be secreted proteins. BIL domains might enhance the variability of secreted proteins by their protein-splicing and cleavage activity as detailed below.

Several, non-exclusive, ways by which BILs function can influence their host proteins are suggested here. BIL activity could be modulated by some external signal. Thus the host protein can be in two states, with and without the BIL domain. These signals might be conformational changes of the host protein or BIL domain (allosteric modification) or change of redox environment in the protein surrounding. Function of BIL domains might not be limited to protein-splicing. They may autocatalyse their cleavage from their host by either N- or C-terminal auto cleavage (Fig. 1). The N-terminal ends of BIL domains, and of Hint domains of inteins and Hog proteins are very similar (Fig. 2). Thus all these domains probably form labile ester bonds on their N-terminal ends. In proteins with BIL domains these ester bonds could be attacked by various nucleophilic molecules, that might include peptides, proteins and small reactive compounds (e.g. glutathione, cysteine). Such reactions would ligate the nucleophiles to a C-terminal position of the host protein and release the BIL and the host protein region downstream to it. This is analogous to Hedgehog protein maturation where the Hint domain mediates the attachment of a cholesterol molecule to the cleaved Hedge domain (Fig. 1). In adhesins with BIL domains this putative ligation might serve to covalently attach the bacteria to its adhesion target. Additionally, released BIL and C-terminal domains could have a function of their own. For example, in pathogenic bacteria that have such proteins, the released domains could serve as decoys to the immune system.

In *Neisseria* strains BILs appear either as short ORFs downstream of MafB genes and in the C-terminal ends of these proteins upstream of a variable domain. This suggests that, at least in *Neisseria*, BILs function as cassettes that can be fused to genes by genetic rearrangement events that may promote the variability of the encoded proteins. Other microevolutionary processes in *Neisseria* and *Ralstonia solanacearum*, a plant pathogen bacterium with a wide host range, are known to generate different C-terminal ends for surface-exposed and virulence proteins (Parkhill *et al.*, 2000; Salanoubat *et al.*, 2002).

Not all species with BILs are pathogens and many pathogenic bacteria with fully sequenced genomes do not have BILs. BILs might be used in processes, not connected with pathogenicity. For example, BIL activity might be one way for bacteria to attach to diverse surfaces.

Conclusions

The bacterial intein-like (BIL) domains we identified appear to have the protein-splicing activity of inteins but we believe their activity serves a different purpose. Whereas inteins are selfish genetic elements we propose that BIL domains contribute to the functionality of the protein in which they reside by protein-splicing and/or autoproteolysis of their host proteins. Our conclusions are based on the types of proteins in which BIL domains reside, the genomic and phylogenetic distribution of BIL domains, and the protein-splicing and autoproteolytic activity of a BIL domain.

Experimental procedures

Data sources

Sequence data was obtained from the NCBI non-redundant sequence databases for *Brucella melitensis* 16 M, *Streptomyces coelicolor* A3(2), *Neisseria meningitidis* Z2491 and *Neisseria meningitidis* MC58 sequences; from the NCBI microbial genome sequences database (http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/genom_table_cgi) for *Pseudomonas syringae* DC3000 (Fouts *et al.*, 2002) (source of preliminary sequence data from The Institute for Genomic Research website at http://www.tigr.org); from Integrated-Genomics (http://www.integratedgenomics.com) for *Rhodobacter capsulatus* SB1003 genome data (Haselkorn *et al.*, 2001); from Joint Genome Institute (http://www.jgi.doe.gov) for the *Rhodobacter sphaeroides* 2.4.1 (Mackenzie *et al.*, 2001), *Magnetospirillum magnetotacticum* MS-1, *Clostridium thermocellum* ATCC 27405 and *Thermobifida fusca* YX genomic sequence data – This data has been provided freely by the US DOE Joint Genome Institute for use in this publication/correspondence only; from The Sanger Institute (http://www.sanger.ac.uk) for the *Neisseria meningitidis* FAM18; from University of Oklahoma, Advanced Center for Genome Technology (http://www.genome.ou.edu) for the *Neisseria gonorrhoeae* FA1090 genome sequence (GenBank accession number for the completed *Neisseria gonorrhoeae* genome is AE004969); and from Baylor College of Medicine Human Genome Sequencing Center (http://www.hgsc.bcm.tmc.edu) for *Mannheimia haemolytica* PHL213 genomic sequence data. BIL domains were named either by host protein name (FhaB and MafB), arbitrarily or by their integratedgenomics database (http://

ergo.integratedgenomics.com/R_capsulatus.html) numbers for *R. capsulatus*, by their Computational Biology Program at ORNL analysis (http://genome.ornl.gov/microbial/rsph) codes for *R. sphaeroides*, by their gene number for *B. melitensis* strain 16 M and *N. meningitidis* strains MC58 and Z2491. Available NCBI gene identifier accessions: SCP1.201 Sco – 13620683, II0519 Bme –17988864, B0369 Nme-B – 7225591, B0372 Nme-B – 7225594, B0655 Nme-B – 7225882, A2115 Nme-A – 15794988, 00588 Rca – 7469167 (more information is provided in *Supplementary material* Tables S1 and S2).

Computational sequence analysis

Sequence searches used the BLAST programs for sequence to sequences searches (Altschul *et al.*, 1997) and the BLIMPS program for blocks to sequences searches (Henikoff *et al.*, 1995). Block multiple sequences alignments were constructed using the BLOCKMAKER (Henikoff *et al.*, 1995) and MACAW (Schuler *et al.*, 1991) programs as described previously (Pietrokovski, 1998). Phylogenetic analysis was done using programs from the PHYLIP package (Felsenstein, 1989) version 3.55.

Functional assay of protein-splicing

In order to create an assay for protein-splicing activity, a plasmid containing the genes for two protein tags was assembled. This plasmid is termed pC2C and is based on the pMALC2 vector (New-England BioLabs (NEB), Beverly, MA). It contains the malE maltose-binding protein (MBP) and the *cbd* gene coding for the chitin-binding domain (CBD) from *B. circulans*. Chitin-binding domain was cloned by PCR from the pTYB2 vector (NEB, Beverly, MA) using the primers: 5′-AAATGTCGACTGCGGTGGCCTGACC-3′ and 5′-TGTCG TATTGCTTCCTTTCGGGCTT-3′. The cloned CBD sequence included the upstream linker 5′-TGCGGTGGCCTGACCG GTCTGAACTCAGGCCTC-3′ and was inserted into the pMALC2 vector between the *Sal*I and *Pst*I restriction sites. The *P. syringae* (Psy) BIL-domain was amplified by PCR from *P. syringae* DC3000 strain genomic DNA (supplied by Dr Sessa G. from the Tel Aviv University, Israel) using the primers 5′-AAAAGGATCCTGCTTTGCGGCCGGAACGA-3′ and 5′-AAAATCTAGAGGTATTATGCACCCATGTCTTG-3′. Polymerase chain reaction (PCR) mixtures containing Taq DNA polymerase (1 μl), Taq polymerase buffer (Sigma, St Louis, MI), 200 mM dNTP, 10 mM of each primer and 100 ng genomic DNA in a 50 μl reaction. Amplification was carried out using a Biometra thermal cycler. BIL-domain was cloned in between the *Bam*HI and *Xba*I sites downstream from the *malE* gene and upstream from the CBD sequence. This pC2C plasmid inserted with the Psy BIL domain is termed pC2C-PsyBIL.

The *R. sphaeroides* (Rsp) BIL2 domain was amplified by PCR from *R. sphaeroides* 2.4.1 strain genomic DNA (supplied by Dr Steven L. Porter, Department of Biochemistry, University of Oxford) using the primers 5′-GAATTCGGTGA TTCATCCTTGGGGCGA-3′ and 5′-TCTAGAAAAACACGGC AAGGGCGAGCGG-3′. BIL-domain was cloned together with 32 amino acids at the N-terminal and 11 amino acids at the C-terminal in between the *Eco*RI and *Xba*I sites downstream from the *malE* gene and upstream from the CBD sequence. This pC2C plasmid inserted with the Rsp BIL2 domain is termed pC2C-RspBIL2.

Expressed fusion proteins originating from pC2C BIL plasmids were termed MBC for MBP-BIL-CBD. pC2C was used as a control plasmid generating the MC fusion protein.

In vitro *protein translation*

In vitro transcription-translation of the proteins MBC and MC by using pC2C BIL plasmids and pC2C plasmid, respectively, as DNA templates was carried out using *E. coli* S30 extract for circular DNA system (Promega, Madison WI). Reaction was carried out using 0.25 mM [^{35}S]-methionine and 220 nmol of plasmid DNA as template following the manufacturer's protocol. Reaction was incubated at 37°C for 90 min for the Psy BIL and 120 min for Rsp BIL. Before electrophoresis, 5 μl or 10 μl of each protein sample were mixed with four volumes of acetone to remove polyethylene glycol from sample. Acetone precipitation was followed by centrifugation at 12 000 *g* for 5 min. Supernatant was discarded and pellet containing the proteins was mixed with protein loading buffer to give a final concentration of 0.06 M Tris-Cl, 2% SDS, 10% v/v glycerol, 0.01% bromophenol blue. Protein were visualized after 10% or 7.5% SDS-PAGE by using a phosphor imaging screen. Signals were then quantified with the NIH Image 1.62 software. Product amounts were from values of three independent experiments averaged for each sample together with their standard deviation of the means. The molar percentage of each product was calculated.

In vivo *expression and purification of Psy BIL*

Competent *E. coli* cells TB1 (NEB, Beverly, MA), were transformed with the pC2C-PsyBIL plasmid described above. The transformed cells were plated on LB agar supplemented with ampicillin (100 μg ml^{-1}). Single colonies were inoculated into 3 ml of LB medium with ampicillin (100 μg ml^{-1}). After incubation for 16 h at 37°C with shaking, 1 ml was used to inoculate a 2 L flask containing 500 ml of LB/Amp100. Incubation was continued at 37°C with shaking until the optical density (OD) at 600 nm was 0.6. Then, IPTG was added to a final concentration of 0.3 mM. After further incubation for 3 h, cells were harvested by centrifugation (5000 *g*, 20 min), resuspended in 20 mM Tris pH 7.4, 200 mM NaCl with a protease inhibitor cocktail (Sigma, St Louis, MI) and lysed by sonication. Lysed cells were then centrifuged at 17000 *g* for 20 min to remove cell debris. Supernatant was then used for all further analysis. Proteins were then affinity purified with either chitin (NEB, Beverly, MA) or amylose beads (NEB, Beverly, MA). Elution of protein from beads before electrophoresis was done by mixing the protein bound beads with SDS-PAGE sample loading buffer.

Antibodies

Western blot analysis was used to identify proteins with either MBP (M) or CBD (C) tags. Monoclonal mouse anti-MBP (Novus Biologicals, Littleton, CO) were used for identification

of the M-tag and polyclonal rabbit anti-CBD (NEB, Beverly, MA) were used for identification of the C-tag. Secondary antibodies used were HRP conjugated goat anti-mouse IgG or goat anti-rabbit IgG (Jackson ImmunoResearch Laboratories, West Grove, PA).

SDS-PAGE and protein staining

The SDS-PAGE was performed as described (Laemmli, 1970). Protein samples were mixed with protein loading buffer to give a final concentration of 0.06 M Tris-Cl, 2% SDS, 10% v/v glycerol, 0.1 M DTT, 0.01% bromophenol blue. All samples were boiled for 3 min before the gel run. TriChromo-Ranger (Pierce, Rockford, IL) prestained markers were used to estimate protein sizes. After electrophoresis, the polyacrylamide gels were fixed in 40% methanol, 7% Acetic acid and then stained by PhastGel Blue R stain (Pharmacia Biotech AB, Sweden). Gel were destained by 40% methanol, 7% acetic acid and then by deionized water. Visualized protein spots were excised using a scalpel before MALDI-TOF analysis.

Electroelution from the gel

Electroelution was performed in GeBAflex – tube (Gene Bio Application, Israel) at 150 V for 2 h. Elution buffer contained 0.025% SDS, Tris and Tricine, pH 8.5. Sodium dodecyl sulphate (SDS) removal after electroelution has been performed by cold TCA:acetone precipitation in the presence of 0.5% sodium deoxycholate (NaDOC) (T. Mehlman and A. Shainskaya, unpublished).

In-gel digestion

Protein bands were excised from the SDS gel stained with PhastGel Blue R stain and destained using multiple washing with 50% acetonitrile in 50 mM ammonium bicarbonate. Protein bands were subsequently reduced, alkylated and in-gel digested with either bovine trypsin (sequencing grade, Roche Diagnostics, Germany) or chymotrypsin (Boehringer Mannheim) applied at a concentration of 12.5 ng μl^{-1} in 50 mM ammonium bicarbonate at 37°C as described (Shevchenko *et al.*, 1996). An extracted peptide solution was dried for subsequent MALDI-MS analysis.

Mass spectrometry

Intact molecular mass measurement and peptide mass mapping were performed on a Bruker Reflex III MALDI time-of-flight (TOF) mass spectrometer (Bruker, Bremen, Germany) equipped with SCOUT source, delayed ion extraction, reflector and a 337 nm nitrogen laser. Each mass spectrum was generated from accumulated data of 200 laser shots. Both external and nearby calibration for proteins were achieved by using BSA and myoglobin proteins, obtained from Sigma. For peptide mapping, internal calibration with molecular ions of regularly occurring matrix ions and peptides derived from trypsin was additionally performed to consolidate further peptide assignment.

Intact molecular weight measurements by MALDI MS

Proteins electroeluted from the gel were further purified by cold acetone precipitation. The dried extract from one lane of the gel was redissolved in 0.5 ml of 80% formic acid and immediately diluted with water to yield a final concentration of 20% formic acid. 50% of this solution was applied to a target plate.

Peptide mass mapping by MALDI MS

Aliquots of one tenth of the extracted peptide mixture volume, dissolved in 0.1% TFA or formic acid/isopropanol/water (1 : 3 : 2) were used for MALDI-MS using fast evaporation or dry droplet methods. The fast evaporation method utilized matrix surfaces made of α-cyano-4-hydroxycinnamic acid (4-HCCA) (Vorm *et al.*, 1994; Jensen *et al.*, 1996). The dry droplet method utilized matrix surfaces made from 2,5-dihydroxybenzoic acid (DHB) (Kussmann *et al.*, 1997).

Acknowledgements

We thank G. Sessa for gift of *P. syringae* bacterial strain, S. L. Porter for gift of *R. sphaeroides* bacterial strain and H. Engelberg-Kulka, G. Amitai and G. Sessa for commenting on the manuscript. Preliminary sequence data was obtained from The Joint Genome Institute (http://www.jgi.doe.gov), The Institute for Genomic Research (http://www.tigr.org), The Sanger Institute (http://www.sanger.ac.uk), University of Oklahoma, Advanced Center for Genome Technology (http://www.genome.ou.edu), IntegratedGenomics (http://www.integratedgenomics.com), and Baylor College of Medicine Human Genome Sequencing Center (http://www.hgsc.bcm.tmc.edu). Sequencing of *P. syringae* DC3000 at TIGR was accomplished with support from NSF: Plant Genome Program. The Gonococcal Genome Sequencing Project supported by USPHS/NIH grant #AI38399, and B.A. Roe, L. Song, S. P. Lin, X. Yuan, S. Clifton, Tom Ducey, Lisa Lewis and D.W. Dyer at the University of Oklahoma. The DNA sequence of *M. haemolytica* PHL213 was supported by grant #00-35204-9229 from USDA/NRICGP to S. Highlander and G. Weinstock at the BCM-HGSC. S. Pietrokovski holds the Ronson and Harris Career Development Chair.

Supplementary material

The following material is available from http://www.blackwellpublishing.com/products/journals/suppmat/mole/mole3283/mmi3283sm.htm

Fig. S1. Assignment of MALDI Peptide Mass to the MC ligation product (Fig. 6A).

A. Sequences detected by MALDI analysis of the MC product are underlined. Twenty-five tryptic peptide masses were assigned to the amino acid sequence of the MC protein, corresponding to sequence coverage of 49%. Amino acids matching the C-tag protein are in italic. The double underlined peptide (ISEFGSTSR-amino acids 388-396) contains the BIL splice site between amino acids Ser393 and Thr394.

B. Measured and calculated masses for tryptic peptides which identify the 50.6 kD MC protein.
Fig. S2. MALDI peptide mapping of the 59.3 kD MB protein (Fig. 6A).
A. Underlined sequences correspond to peptides detected by MALDI. Uppercase letters match amino acids of the M-tag and lowercase letters match those of the BIL domain. Note that the C-terminus of the protein, Asn 541, is the penultimate C-terminal residue of the BIL sequence (Fig. 4).
B. Measured and calculated molecular mmasses of the two C-terminal peptides.
Table S1. BIL sequence motifs.
Table S2. BIL sequence sources.

References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389–3402.

Aspock, G., Kagoshima, H., Niklaus, G., and Burglin, T.R. (1999) *Caenorhabditis elegans* has scores of hedgehog-related genes: sequence and expression analysis. *Genome Res* **9:** 909–923.

Belfort, M., and Roberts, R.J. (1997) Homing endonucleases: keeping the house in order. *Nucleic Acids Res* **25:** 3379–3388.

Coote, J.G. (1992) Structural and functional relationships among the RTX toxin determinants of gram-negative bacteria. *FEMS Microbiol Rev* **8:** 137–161.

Dalgaard, J.Z., Moser, M.J., Hughey, R., and Mian, I.S. (1997) Statistical modeling, phylogenetic analysis and structure prediction of a protein splicing domain common to inteins and hedgehog proteins. *J Comput Biol* **4:** 193–214.

Felsenstein, J. (1989) *PHYLIP – Phylogeny Inference Package* (Version 3.2). Cladistics 5: 164–166.

Fouts, D.E., Abramovitch, R.B., Alfano, J.R., Baldo, A.M., Buell, C.R., Cartinhour, S., *et al.* (2002) Genomewide identification of *Pseudomonas syringae* pv. tomato DC3000 promoters controlled by the HrpL alternative sigma factor. *Proc Natl Acad Sci USA* **99:** 2275–2280.

Hall, T.M., Porter, J.A., Young, K.E., Koonin, E.V., Beachy, P.A., and Leahy, D.J. (1997) Crystal structure of a hedgehog autoprocessing domain: homology between hedgehog and self-splicing proteins. *Cell* **91:** 85–97.

Haselkorn, R., Lapidus, A., Kogan, Y., Vlcek, C., Paces, J., Paces, V., *et al.* (2001) The Rhodobacter capsulatus genome. *Photosynthesis Res* **70:** 43–52.

Henikoff, S., Henikoff, J.G., Alford, W.J., and Pietrokovski, S. (1995) Automated construction and graphical presentation of protein blocks from unaligned sequences. *Gene* **163:** 17–26.

James, R., Kleanthous, C., and Moore, G.R. (1996) The biology of E colicins: paradigms and paradoxes. *Microbiology* **142:** 1569–1580.

Jensen, O.N., Podtelejnikov, A., and Mann, M. (1996) Delayed extraction improves specificity in database searches by matrix-assisted laser desorption/ionization peptide maps. *Rapid Comms Mass Spectrometry* **10:** 1371–1378.

Kussmann, K., Nordhoff, E., Rahbek-Nielsen, H., Haebel, S., Rossel-Larsen, M., Jakobsen, L., *et al.* (1997) Matrix-assisted laser desorption/ionization mass spectrometry sample preparation techniques designed for various peptide and protein analytes. *J Mass Spectrometry* **32:** 593–601.

Laemmli, U.K. (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227:** 680–685.

Mackenzie, C., Choudhary, M., Larimer, F.W., Predki, P.F., Stilwagen, S., Armitage, J.P., *et al.* (2001) The home stretch, a first analysis of the nearly completed genome of *Rhodobacter sphaeroides* 2.4.1. *Photosynthesis Res* **70:** 19–41.

Naumann, M., Rudel, T., and Meyer, T.F. (1999) Host cell interactions and signalling with *Neisseria gonorrhoeae*. *Curr Opin Microbiol* **2:** 62–70.

Parkhill, J., Achtman, M., James, K.D., Bentley, S.D., Churcher, C., Klee, S.R., *et al.* (2000) Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. *Nature* **404:** 502–506.

Paruchuri, D.K., Seifert, H.S., Ajioka, R.S., Karlsson, K.A., and So, M. (1990) Identification and characterization of a *Neisseria gonorrhoeae* gene encoding a glycolipid-binding adhesin. *Proc Natl Acad Sci USA* **87:** 333–337.

Paulus, H. (2000) Protein splicing and related forms of protein autoprocessing. *Annu Rev Biochem* **69:** 447–496.

Pietrokovski, S. (1994) Conserved sequence features of inteins (protein introns) and their use in identifying new inteins and related proteins. *Protein Sci* **3:** 2340–2350.

Pietrokovski, S. (1998) Modular organization of inteins and C-terminal autocatalytic domains. *Protein Sci* **7:** 64–71.

Pietrokovski, S. (2001) Intein spread and extinction in evolution. *Trends Genet* **17:** 465–472.

Porter, J.A., Ekker, S.C., Park, W.J., von Kessler, D.P., Young, K.E., Chen, C.H., *et al.*(1996a) Hedgehog patterning activity: role of a lipophilic modification mediated by the carboxy-terminal autoprocessing domain. *Cell* **86:** 21–34.

Porter, J.A., Young, K.E., and Beachy, P.A. (1996b) Cholesterol modification of hedgehog signaling proteins in animal development. *Science* **274:** 255–259.

Salanoubat, M., Genin, S., Artiguenave, F., Gouzy, J., Mangenot, S., Arlat, M., *et al.* (2002) Genome sequence of the plant pathogen *Ralstonia solanacearum*. *Nature* **415:** 497–502.

Schuler, G.D., Altschul, S.F., and Lipman, D.J. (1991) A workbench for multiple alignment construction and analysis. *Proteins: Structure, Function, Genetics* **9:** 180–190.

Shevchenko, A., Wilm, M., Vorm, O., and Mann, M. (1996) Mass spectrometric sequencing of proteins from silver-stained polyacrylamide gels. *Anal Chem* **68:** 850–858.

Smith, A.M., Guzman, C.A., and Walker, M.J. (2001) The virulence factors of Bordetella pertussis: a matter of control. *FEMS Microbiol Rev* **25:** 309–333.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple

sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acid Res* **22:** 4673–4680.

Vorm, O., Roepstorff, P., and Mann, M. (1994) Improved resolution and very high sensitivity in MALDI-TOF of matrix surfaces made by fast evaporation. *Anal Chem* **66:** 3281–3287.

Xu, M.-Q., and Perler, F.B. (1996) The mechanism of protein splicing and its modulation by mutation. *EMBO J* **15:** 5146–5153.

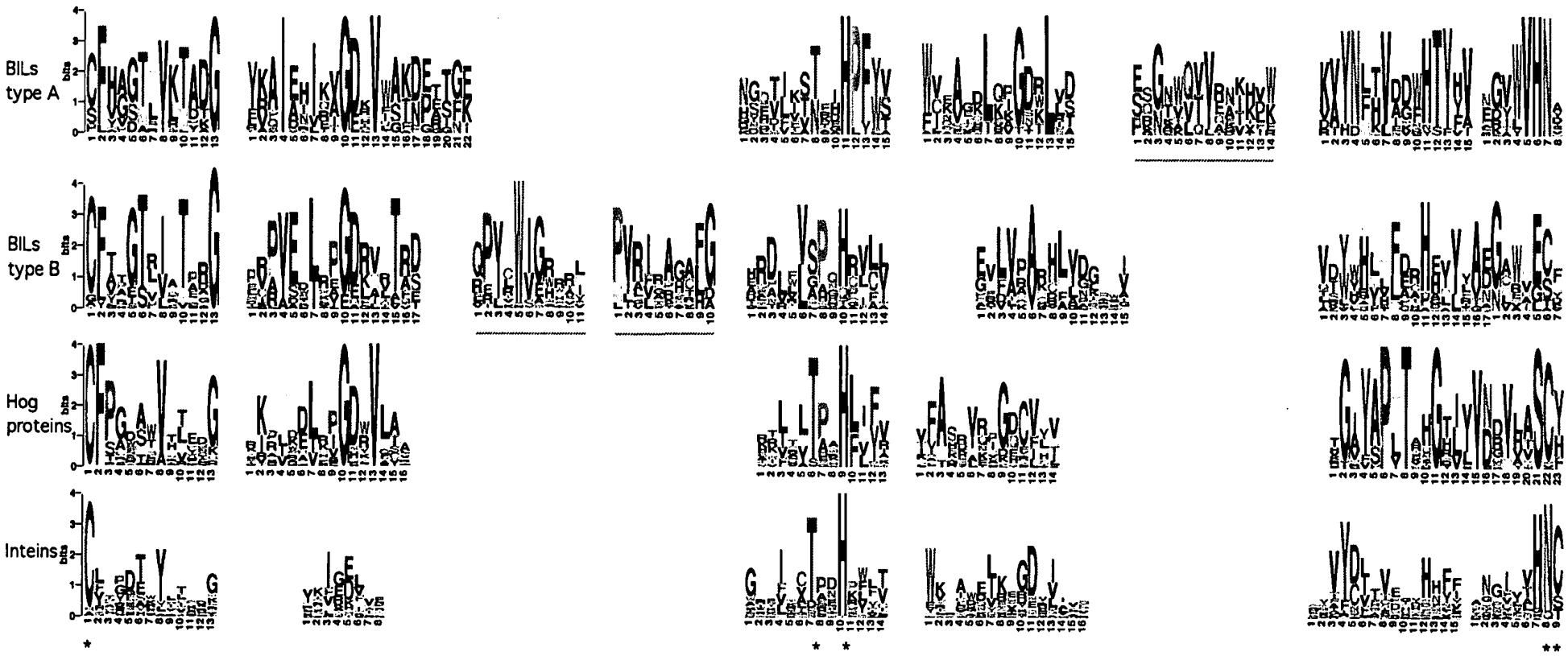

Fig. 2 (high resolution). Conserved motifs of Hint protein domains. Each row shows conserved motifs from one type of Hint protein domain. Motifs are ordered left to right in the N' to C' positions along the protein sequences. Similar motifs are vertically aligned with each other. Unique A-type and B-type BIL motifs are underlined with hatched lines. The motifs are shown as sequence logos where the height of amino acids are proportional to their conservation in each position. Positions of the intein protein-splicing active site residues are marked by asterisks. Protein motifs were found and are displayed as previously described (Pietrokovski, 1998). The BIL motif sequences and the distances between consecutive motifs are listed in supplementary Table 1. Intein and hedgehog Hint domains are those described by Aspock (1999) and Pietrokovski (2001). Only intein and hedgehog motifs common to Hint domains are shown.

Table S1. BIL sequence motifs.[a]

A-type

A2115+_neime	9	SFHGSTLVKTADG	(0)	YKAIAHIQAGDRVFAKDETSGK	(27)	NNQTLISNKIHPFYS	(3)	WIQAGRLKKGDTLLS	(0)	ESGAKQTVQNITLK	(4)	KAYNLTVADWHTYFV	(7)	EGVWVHNE	151
B0369+_neime	11					NSQILISNRIHPFYS	(3)	WIKAEDLKAGSRLLS	(0)	ESGRTQTVRKTVVK	(4)	KAYNLTVADWHTYFV	(7)	EGVWVHNS	91
B0372+_neime	6	SFHGSTLVKTADG	(0)	YKAIARIRTGDRVFAKDEASGK	(27)	NNQTLISNKIHPFYS	(3)	WIQAGRLKKGDTLLS	(0)	ESGAKQTVQNITLK	(4)	KAYNLTVADWHTYFV	(7)	EGVWVHND	148
B0655+_neime	9	SFHGSTLVKTADG	(0)	YKAIARIRTGDRVFAKDEASGK	(27)	NNQTLISNKIHPFYS	(3)	WIQAGRLKKGDTLLS	(0)	ESGAKQTVQNITFK	(4)	KAYNLTVADWHTYFV	(7)	EGVWVHND	151
MafB1_neimeC	348	SFHGSTLVKTADG	(0)	YKAIAHIRVGESVFAKDETSGK	(27)	NSQTLISNRIHPFYS	(3)	WIQAGRLKKGDTLLS	(0)	ESGAKQTVQNITLK	(4)	KAYNLTVADWHTYFV	(7)	EGVWVHND	490
BIL2_neimeC	9	PLYVGALVKTADG	(0)	YKAIAHIRVGESVLSKDEASGK	(27)	NSQTLISNRIHPFYS	(3)	WIQAGRLKKGDTLLS	(0)	ESGAKQTVQNITFK	(4)	KAYNLTVADWHTYFV	(7)	EGVWVHNA	151
BIL3_neimeC	11					NSQILISNRIHPFYS	(3)	WIKAEDLKAGSRLLS	(0)	ESGKTQTVRNIVVK	(4)	KAYNLTVADWHTYFV	(7)	EGVWVHND	91
BIL4_neimeC	9	SFHGSTLVKTADG	(0)	YKAIAHIRVGESVLSKDEASGK	(27)	NSQTLVSNKIHPFYS	(3)	WIKAEDLKAGSRLLS	(0)	ESGKTQTVRNIVVK	(4)	KAYNLTVADWHTYFV	(7)	EGVWVHNA	151
BIL5_neimeC	9	SFHGSTLVKTADG	(0)	YKAIARIRTGDRVFAKDEASGK	(27)	NNQTLISNKIHPFYS	(3)	WIQAGRLKKGDTLLS	(0)	ESGAKQTVQNITFK	(4)	KAYNLTVADWHTYFV	(7)	EGVWVHNA	151
BIL6_neimeC	9	SFHGSTLVKTADG	(0)	YKAIAHIQAGDRVLSKDEASGE	(27)	NSQTLVSNKIHPFYS	(3)	WIQAGRLKKGDTLLS	(0)	ESGAKQTVQNITLK	(4)	KAYNLTVADWHTYFV	(7)	EGVWVHNS	151
MafB1_neigo	347	SFHGSTLVKTADG	(0)	YKAIAHIQAGDRVLSKDEASGE	(27)	NSQTLISNRIHPFYS	(3)	WIKAEDLKAGSRLLS	(0)	ESGKTQTVRNIVVK	(4)	KAYNLTVADWHTYFV	(7)	EGVWVHND	489
BIL2_neigo	9	PFHGSTLVKTADG	(0)	YKAIARIRVGDHVFAKDEASGE	(27)	NNQTLISNRIHPFYS	(3)	WIKAEDLKAGSRLLS	(0)	ESGRTQTVRNIIVK	(4)	KAYNLTVADWHTYFV	(7)	EGVWVHNA	151
BIL3_neigo	9	SFHGSTLVRTADG	(0)	YKAIAHIQAGDRVLSKDEASGK	(27)	NSQTLISNRIHPFYS	(3)	WIKAEDLKAGNRLFA	(0)	ESGKTQTVRNIVVK	(4)	KAYNLTVADWHTYFV	(7)	EGVWVHNS	151
MafB2_neigo	347	SFHGSTLVKTADG	(0)	YKAIAHIQAGDRVLSKDEASGE	(27)	NSQTLISNRIHPFYS	(3)	WIKAEDLKAGSRLLS	(0)	ESGKTQTVRNIVVK	(4)	KAYNLTVADWHTYFV	(7)	EGVWVHND	489
BIL5_neigo	9	SFHGSTLVKTADG	(0)	YKAIAHIQAGDRVLSKDEASGE	(27)	NSQTLISNRIHPFYS	(3)	WIKAEDLKAGSRLFA	(0)	ESGKTQTVRNIIVK	(4)	KAYNLTVADWHTYFV	(7)	EGVWVHND	151
BIL6_neigo	9	PFHGSTLVKTADG	(0)	YKAIAHIQTGEHVFAKDETSGK	(27)	NSQTLISNRIHPFYS	(3)	WIKAEDLKAGSRLLS	(0)	ESGRTQTVRNTVVK	(4)	KAYNLTVADWHTYFV	(7)	EGVWVHNS	151
3875_87_magma	292	CFVAGTPVRMADG	(1)	EKAIETVEIGEQVQGTDGTINE	(17)	NSLDFFVTADHPFLT	(17)	ALNVTQLVIGDTLIT			(19)	VVYNLHLIGNNTYVA	(0)	SGYYVHNY	433
FhaB_psesy	5987	CFAAGTMVSTPDG	(0)	ERAIDTLKVGDIVWSKPEGGGK	(31)	EDESLLVTPGHPFYV	(5)	FVPVIDLKPGDRLQS			(23)	KTYNLTVDVGHTFYV	(2)	LKTWVHNT	6135
BIL1_strco	1082	SFPAGTRVLMADG	(1)	RRSIEQIEAGDLVTATDPTTGE	(24)	DGSTLTSTTHHPYWS	(5)	WKNAGDLEAGDTLRT	(0)	PQNTAVVIAATHDW	(4)	DAYDLTVDGFHSYYV	(4)	TDVLVHNN	1221
39_9_thefus	299	SFVPGTLVLLADG	(1)	YAPIETITVGDDVWAFDPRTGT	(28)	HGGVVVATDAHPFWV	(5)	WVAAIDLEPGTWLRT	(0)	SAGTWVQVRAVAVR	(5)	RVHNLTVADLHTYYV	(4)	ADALVHNE	443
FhaB_manha	2865	SFHGDMEVKTDKG	(0)	YRQISSIKVGDKVLAKNERTGI	(27)	KYHTIVSNKIHPFFT	(24)	WVDAQHLQKGYRLLA	(0)	ESGEWQTVTKVKIK	(4)	KAYNMTVEKDHTYFI	(7)	EGVWVHND	3028
BIL1_cloth	48	CFVAGTLILTVAG	(0)	LVAIENIKAGDKVIATNLETFE	(22)	NGEVIKTTFEHPFYV	(4)	FVEAKELQVGDKLLD	(0)	SKGNVLVVEEKKLE	(6)	KVYNFKVDDFHTYHV	(2)	NGILVHNA	183
BIL2_cloth	31	CFVAGTMILTATG	(0)	LVAIENIKAGDKVIATNPETFE	(22)	GXEVIKTTLGHLFYV	(4)	FVEAVKLQPTDKLVD	(0)	SGGNVLVVEXKKFE	(6)	KVYNFKVNDFYTYHV	(2)	NGILVHNV	166
BIL3_cloth	70	CFVAGTMILTATG	(0)	LVAIENIKAGDKVIATNPETFE	(22)	NGEVIKTTFEHPFYV	(4)	FVEAGKLQIGDRLVD							167
BIL4_cloth	316	CFVAGTMILTATG	(0)	LVAIENIKAGDKVIATNPETFE	(22)	GGEVIKTTFDHPFYV	(4)	FVEAGKLQVGDKLLD	(0)	SRGNVLVVEEKKLE	(6)	KVYNFKVDDFHTYHV	(2)	NEVLVHNA	451
BIL5_cloth	31	CFVAGTMILTTTG	(0)	LVAIENIKAGDKVIATNPETFE	(22)	GGEVIKTTFDHPFYV	(4)	FVEAKQLHVGDKLLD	(0)	SKGNVLVVEDKKIK	(6)	KVYNFQVADFHTYHV	(2)	NGVLVHNV	166
BIL6_cloth	87	CFVAGTMILTVAG	(0)	LVAIENIKAGDKVIATNPETFE	(22)	NGDVIKTTFEHLFYA	(4)	FVEAKELQVGDKLLD	(0)	SKGNVLVVEDKKIK	(6)	KVYNFQVDDFHTYHV	(2)	NGVLVHNV	222
BIL7_cloth	17							FVKEMKLQPGNRLVD	(0)	SKGNVLVVEEKKLE	(6)	KVYNFKVNDFHTYHV	(2)	DGILVHNA	76
BIL8_cloth	15	CFVAGTMILTATG	(0)	LVAIENIKAGDKVIATNPETFE	(22)	GXEIIKTTLGHLFYV	(4)	FVEAVKLQPTDKLVD	(0)	SGGNVLVVEXKKFE	(6)	KVYNFKVNDFYTYHV	(2)	NGILVHNV	143
BIL9_cloth	8	CFVAGTMILTATG	(0)	LVAIENIKAGDKVIATNPETFE	(22)	GXEIIKTTLGHLFYV	(4)	FVEAVKLQPTDKLVD	(0)	SGGNVLVVEXKKFE	(6)	KVYNFKVNDFYTYHV	(2)	NGILVHNV	157
BIL10_cloth	1											VYNFKVDNFHTYHV	(2)	NRVLVHNA	24

B-type

00126_rhoca	160	CFTPGTLIDTPAG	(0)	PRPVEALRPGDRVSTRD	(3)	QEILWIGSRRM	(12)	PVRLGAVRLG	(14)	AADLLVSPQHRVLV	(12)	EVLVQACDLVDDAAV	(8)	VTYLHLLFARHQVIRAN	(0)	GVETESF	312
00199_rhoca	37	GFYGETVLQTARG	(0)	LRRVSSILEGEKMRTFT	(3)	APVLSIERFAL	(12)	PLSLPAGLFG	(1)	TRNRFVAPEQCLLL	(12)	LLLVPAKVLGLLPQV	(8)	AVLYRLLFERPELVVTD	(1)	GAVMLCD	177
00459_rhoca	38	GFAAGTRVRTPAG	(0)	LRRIETLKPGDLVETQE	(3)	QPVVAVERTRL	(7)	PIRFAAGAHG	(1)	ERPVLVAPQQRVLV	(12)	EVLVAARTLVDGEMV	(7)	VDYVRLVFDCAHMVFAE	(0)	GLAVECF	171
00460_rhoca	174	CFAPSTPIATPGG	(0)	DCPAASLKAGDLVLTAD	(3)	QPILWSGRIAL	(7)	PVRLCAPAFG	(1)	TRDLWVLPQHRVAL	(12)	EVLVPAHHLVDGISA	(8)	LSWHGLLLQGHHLLIAD	(0)	GCRVESL	308
00588_rhoca	443	CFTAGTLIETPRG	(0)	PVPVESLRAGDLVVTRD	(3)	VPVLWSGGRSL	(12)	PVAIRENALG	(1)	HGALLLSPQHAVLA	(5)	ERLVRARHLAGLNDP	(10)	VSYHHILLERHGIVTAN	(0)	GLACESL	577
00746_rhoca	85	ALARGSVLMTEDG	(0)	PVAIEDLQPGQGVLTAE	(3)	ERVCWIGSMVI	(15)	LTRITAEAFG	(4)	ALDLVLGPRARLCL	(12)	AADVPARAFLDGISV	(8)	VTVYHVVLEQHGSLRVA	(0)	GLEVEAF	230
00949_rhoca	125	CLGTGTMIATAEG	(0)	PAPIDWLRPGDRVLTRD	(3)	QPLLWVGQHTM	(9)	PLLLSAACFG	(4)	ERDVLLSPGTGVLL	(12)	EMFAKARHALPKAEA	(4)	QKLYSMLLATPEVVLAE	(0)	GMWVGSV	260
01216_rhoca	128	CFAAGTLIATRRG	(0)	PKPVEDLGPEDRLQTSD	(3)	RPVQWVGRWRV	(7)	PVRFAPGVLG	(1)	DRALFLSGQHRVLI	(10)	EVLVAAKALVGLPGI	(7)	VDWVHVMMPTHEVIFAE	(0)	NARAETM	259
01374_rhoca	144	AFTTGTLITMAGG	(1)	QRPIETLAPGDRVLTRD	(3)	QPVRLVARATL	(7)	PVVISAGTLG	(1)	ESDLVVAPHHRVFL	(12)	EILVQAKHLVDGEHV	(7)	VDYFALVFDRHEIVYAE	(0)	GVPVESL	278
01522_rhoca	1872	CFTPGTLIATPKG	(0)	ERLVEELREGDKILTRD	(3)	QEIRWIGRTDL	(12)	PVLIRAGSLG	(4)	ERDMLVSPNHRMLV	(12)	EVLVAAKHLIDNRGV	(7)	TSYIHFMFDRHEVVLGN	(0)	GAWTESF	2013
01523_rhoca	85	CFTATSLIATGQG	(0)	GVPVSELVPGARVITRD	(3)	QELLWVGRRRF	(12)	PVRIAAGALG	(4)	ERDMLVSPNHRFLT	(9)	ERLTMARDLVGLDGI	(7)	VDYWQLLFAHHELVLAD	(0)	GAWSESF	223
01524_rhoca	89	CLTPGTLIETKRG	(0)	QVPVEKLRPGDRVLTRD	(3)	QPIRWIGRRRL	(12)	PVRIAAGALG	(4)	ETDMLVSPQHRMLI	(12)	EVLAAALHMLGQPGI	(7)	VTYLHLMLDAHEIIRAN	(0)	GAWTESF	230
02710_rhoca	541	CLVAGSRVSTPRG	(0)	PVPVEDLRPEDLVTVRD	(3)	LPVLWIGRRRV	(12)	PVEIGAGRLG	(1)	AAPVRLSALHGIAV	(2)	GFLARAGHLAATGWG	(14)	VLYLHLLLPRHALLSVE	(0)	GLWVESF	676
03530_rhoca	39	GFAMGSRVATMDG	(0)	LLPVEFLNLGDRIVTRS			(16)	LVGIAPGALG	(4)	GQAMVLGSGTQVLL	(12)	QALVAVERLIDGQFI	(7)	IRIFALHFEAPEVIYAD	(0)	GVEIGCK	170
4825_rhosp	15	CFTPGTLIATVRG	(0)	EVAVEALAAGDRIVTRD	(3)	QPLRWISRRRL	(12)	PVLIEKGSLG	(4)	DRDMMVSPNHRILV	(12)	EVLVAAKHLVGPRGI	(7)	TTYLHLMFDRHEVVLAN	(0)	GAWTESF	156
BIL2_rhosp	34	SLTAGTPVLTLAG	(0)	IRPAEGIRPGDRLVARS			(17)	MVAIGASTLA	(4)	DETLLVPADQPLLL	(12)	PVVLPARRLVDGQLT	(7)	VDLVTLTFAAPAAIYAS	(0)	ELHPVTR	167
BIL1_brusu	86	CLLKGTLVTTPNG	(0)	PVAVEKLCVGDLVTTVS	(3)	LPIKWIGWQNY	(12)	PIRVRRHALD	(4)	HRDLYLSPNHALFI	(1)	GVLIRVKDLVNGRSI	(8)	LDYYNIVLDRHAVVLAE	(0)	GAAVETF	217
BIL2_magma	126	CFVTGTMIATARG	(0)	EVAVEDLRAGDFARTAE	(3)	RPIVWIGHREI	(12)	PVRVRTGAFG	(4)	ARDLYLSPGHPVLV	(8)	GTLVPI					223
BIL3_magma	115	CYVTGTRIRTERG	(0)	EIAVEDLQVGDFAVTAS	(3)	RPITWIGHREI	(12)	PVRVRAGAFG	(4)	VNDLFLSPGHPVLV	(8)	GVLVPVMCLINGTTI	(7)	VTYWHVELDAHDILLAE	(0)	GLPAESY	252
BIL4_magma	116	CFVSGTRISVERG	(0)	SIPVELLRIGEKARLAS	(3)	RTITWIGHREI	(12)	PVRVRAGAFG	(4)	ARDLFLSPGHPVLI	(8)	GVLVPVMCLINGTSI	(7)	VTYWHVELDRHDILLAE	(0)	GLPAESY	253

a. The N- and C-terminal positions of the BIL domain in its host protein are listed before and after the BIL motifs, respectively. In parenthesis between motifs are the number of not shown intervening residues. Some domains are partial, missing N- or C-parts.

Table S2. BIL sequences sources.[a]

Name	Source	Date	Contig/Entry	Coordinates
39_9 Tfu	JGI	1Nov00	39	13655-15508
SCP1.201 Sco	NCBI		13620683 +32 N' aa	
3875_87 Mma	NCBI		21614488	76532-75165
B0369+ Nme-B	NCBI		7225591 +34 N' aa	
B0372+ Nme-B	NCBI		7225594 +11 N' aa	
B0655+ Nme-B	NCBI		7225882 +14 N' aa	
A2115+ Nme-A	NCBI		15794988 +24 N' aa	
MafB1 Nme-C	Sanger	15May02	NmC	1833717-1835480
BIL2 Nme-C	Sanger	15May02	NmC	1836857-1837573
BIL3 Nme-C	Sanger	15May02	NmC	1838418-1838981
BIL4 Nme-C	Sanger	15May02	NmC	1839771-1840439
BIL5 Nme-C	Sanger	15May02	NmC	627204-627920
BIL6 Nme-C	Sanger	15May02	NmC	628395-629102
MafB1 Ngo	OU-ACGT	26Sep00	AE004969	1560214-1561941
BIL2 Ngo	OU-ACGT	26Sep00	AE004969	1563413-1564129
BIL3 Ngo	OU-ACGT	26Sep00	AE004969	1565033-1565809
MafB2 Ngo	OU-ACGT	26Sep00	AE004969	1355876-1354062
BIL5 Ngo	OU-ACGT	26Sep00	AE004969	1351509-1350766
BIL6 Ngo	OU-ACGT	26Sep00	AE004969	1349978-1349310
FhaB Psy	TIGR	30Aug02	5668	5148986-5149429
FhaB Mha	BCM	4Oct01	C78-C79-C80-C81 C82-C83-C84-C85	11046-20977
BIL1 Cth	NCBI		22262155	5528-4683
BIL2 Cth	NCBI		22262016	2476-1667
BIL3 Cth	NCBI		22262092	2185-2685
BIL4 Cth	NCBI		22262016	7936-6224
BIL5 Cth	NCBI		22262016	4623-3736
BIL6 Cth	NCBI		22262092	1-1035
BIL7 Cth	NCBI		22262145	1412-2059
BIL8 Cth	NCBI		22262205	1202-462
BIL9 Cth	NCBI		22262260	657-1
BIL10 Cth	NCBI		22262017	34594-34986
BIL11 Cth	NCBI		22262176	416-728
4825 Rsp	JGI	26Mar01	184	67785-67165
BIL2 Rsp	JGI	26Mar01	177	9673-10194
00588 Rca	NCBI		3128319	
01522 Rca	IG	Dec01	1A01-1C09	273248-279442
02710 Rca	IG	Dec01	1D09-1F02	197288-199588
01524 Rca	IG	Dec01	1A01-1C09	280569-281423
01523 Rca	IG	Dec01	1A01-1C09	279638-280444
00126 Rca	IG	Dec01	2G06-2D11	114767-113670
01216 Rca	IG	Dec01	2A12-2D05	222590-223555
00949 Rca	IG	Dec01	2A12-2D05	325707-326651
01374 Rca	IG	Dec01	2A12-2D05	148470-149462
00459 Rca	IG	Dec01	2G06-2D11	434469-435083
00460 Rca	IG	Dec01	2G06-2D11	435094-436191
00746 Rca	IG	Dec01	2D10-2D06	2243-3079
03530 Rca	IG	Dec01	1A01-1C09	521700-521173
00199 Rca	IG	Dec01	2G06-2D11	178648-177806
BIL2 Mma	NCBI		21613062	922-1590
BIL3 Mma	NCBI		21614112	2449-1475
BIL4 Mma	NCBI		21614173	2216-3187

BIL5 Mma	NCBI	21612572	3-338
BIL6 Mma	NCBI	21613847	2033-1774
II0519 Bme	NCBI	17988864	

a. The sources are named as follows JGI - Joint Genome Institute (http://www.jgi.doe.gov), NCBI - (http://www.ncbi.nlm.nih.gov), Sanger - The Sanger Institute (http://www.sanger.ac.uk), OU-ACGT University of Oklahoma, Advanced Center for Genome Technology (http://www.genome.ou.edu), TIGF Institute for Genomic Research (http://www.tigr.org), BCM - Baylor College of Medicine Human Gen Sequencing Center (http://www.hgsc.bcm.tmc.edu) and IG - IntegratedGenomics (http://www.integratedgenomics.com).

Dates refer to the data release dates used. The NCBI entries were extended as noted. Coordinates o BIL host protein ORF are given for nucleotide contigs/entries. The positions of BILs within these O listed in Table S1.